Technische Universität München

Department of Mathematics

Modeling the rental price per square meter in major German cities

Master's Thesis

by

Ugochukwu Onumadu

Supervisor:	Prof. Ph.D. Claudia Czado
Advisor:	Prof. Ph.D. Claudia Czado, M.Sc. Hassan H. Alnasser
Submission Date:	24.03.2022

With my signature below, I assert that the work in this thesis has been composed by myself independently and no source materials or aids other than those mentioned in the thesis have been used.

München, 24. März 2022

Place, Date

Signature

Acknowledgement

I would like to express my profound gratitude to my able supervisor and advisor, Prof. Ph.D. Claudia Czado for her tremendous efforts, recommendations and advice, which guided me throughout this research. Honestly, I learned so much under her supervision.

I also express my sincere gratitude to my advisor M.Sc. Hassan H. Alnasser, for his outstanding advice, encouragement, and support throughout this research work.

I will equally remain grateful to all my lecturers at TUM, especially in the Mathematics Department. I highly appreciate the kindness of the FDZ Ruhr at RWI (and ImmobilienScout24) institution for providing the data used in this thesis.

I will not fail to commend the efforts of my lovely beautiful wife (Divine), my lovely children (Prince, Jessica, and John) whom I have not seen for many years now for their patience, encouragements, and support throughout my studies at TUM.

I will equally extend my appreciation to my brothers and sisters, Precious Seed family, CREM family, my friends and well-wishers for their prayers and support throughout this research work.

Above all, I give thanks to Almighty God for His enablement in my studies at TUM. To God be the Glory.

Abstract

This thesis considers a statistical model for the apartment rent price per square meter (**rent_sqm**) in Germany. The topic of apartment rent prices in Germany became relevant in major German cities since there is a strong increase in rent prices and in particular, the development of new apartment buildings is lacking. A data, collected by the FDZ Ruhr at RWI (and ImmobilienScout24) institution, is analyzed in this research. The data consists of 2.6 million apartments and 59 variables. We focus on the most relevant 31 variables such as "the additional cost", "heat cost", "living space", etc., and the cities **Munich and Berlin** for two time periods: **2015 and 2019**. This will enable us to compare the behaviour of **rent_sqm** prices in both cities at two different time periods. Once we have done an exploratory data analysis to identify the significant covariates and useful model formulation, we decide to fit a multiple linear regression model (LM). The final fitted model also includes interaction terms between the variables. To assess model fit, we use the adjusted multiple coefficient of determination (R^2_{adj}) and the analysis of variance (ANOVA) ratio test to measure the fit and the complexity of the model, respectively. After we have found a suitable model, we use it for prediction within different scenarios. The results show that rent price per square meter is exponentially increasing over time in Munich and Berlin. Further, Munich has higher rent prices than Berlin. When **a pet is allowed**, it **decreases** the rent price per square meter in both cities, while **Upscale furnishing** apartments as well as apartments with **a parking space** increase rent_sqm in both cities. In Berlin, rent_sqm increases with respect to the order of the **energy efficiency categories (Low, Medium, and High)** as well as the order of the **number of bedrooms (0-1, 2,>2)** in both time periods. However, in Munich, rent_sqm decreases in this order. Furthermore, the rent price per square meter decreases when renting larger apartments in Munich, whereas, in Berlin, this is not the case. Considering the predictions with our four scenarios: scenario 1 (smaller), scenario 2 (small), scenario 3 (large) and scenario 4 (larger) apartments, the **rent_sqm** increased from 2015 to 2019 in Berlin by 8.70%, 38.16%, 29.16%, and 69.47% for smaller, small, large, and larger apartments. In Munich, on the other hand, **rent_sqm** increased in the same period of time by 1.84%, 9.27%, 70.71%, and 60.13%, respectively. Also, in Berlin 2015, the **rent_sqm increased** from scenario 1 to scenario 2, scenario 1 to scenario 3, and scenario 1 to scenario 4 by 36.31%, 54.35% and 67.73%, respectively, while ,in Munich 2015, **rent_sqm decreased** by 8.11%, 10.69% and 18.66%, respectively. Munich 2019, however, shows a different trend. The **rent_sqm decreased** by 1.40% from scenario 1 to scenario 2 and **increased** by 49.70% and 27.89% from scenario 1 to scenario 3 and scenario 1 to scenario 4, respectively.

Contents

Chapter 1

Introduction

The importance of using statistical methods to develop a mathematical equation that models the relationship between a response variable **rent_sqm** and a set of explanatory variables can not be over-emphasized as the demand for renting an apartment in Germany, especially in Munich and Berlin is relatively high compared to other cities. I experienced this situation on my arrival in Munich city of Germany as I found it difficult in getting a place to settle down for my studies. This triggered my curiosity to investigate what could be the cause as between 2011 and 2016, about 45,000 new apartment were built in Munich for roughly 90,000 people even as the population in Munich rose from 200,000 to 1.55 million during the same period of time (Mobert, 2017). Therefore, about 55,000 more apartments were needed to accommodate the new arrivals and by 2030, about 150,000 apartments would be needed as the population will increase to more than 1.7 million based on the estimate of Mobert (2017). Germany is representative of the situation in many aspects when compared to other high-income countries like the UK, France, the US, Canada, etc, therefore, apartment prices and rents are causing serious problems as they have risen extremely in the countries' large cities (Lutz, 2020). In international comparisons, like Northern America or Southern Europe, Germany has a higher share of renters. For instance, in 2018, the homeownership rate in Germany was 51.5% compared to 65.1%, 72.4%, and 96.4% in the UK, Italy, and Romania respectively (Lutz, 2020).

When it comes to property, Munich is the most active city in Germany with its fast-rising population and historically low vacancy rate. This may lead to a further price increase. Also in Berlin, further price rises are equally expected (Mobert, 2017).

Between 2009 and 2017, apartment prices in Munich have doubled, and within the same period of time, the population in Munich rose from 1.36 million to 1.53 million. In Berlin, however, house and apartment prices increased by 10% from 2017 to 2018 (Möbert et al., 2018). Munich, Berlin, Frankfurt, etc, will still see the highest increase in property prices and rents (Kholodilin and Mense, 2012).

Gustafsson and Wogenius (2014) investigated the factors that are of most statistical sig-

nificance for the sales prices of apartments in the Stockholm City Centre using multiple linear regression. They found that proximity to public transport is a driver for the price of an apartment.

Also, Thomschke (2015) carried out research work on the changes in the distribution of rental prices in Berlin from 2007 to 2012. He used a quantile regression model and decomposition methods for his analysis. He found that rent price increased throughout the entire distribution, but more at higher quantiles.

In this thesis, our goal is to investigate the relationship between rent per square meter charged for an apartment characterized by a set of continuous and discrete covariates. We will investigate whether we need a transformation of the response and whether covariates enter the model and whether interaction terms are required. We focus on Berlin and Munich for the time points 2015 and 2019. We use multiple linear regression to fit our model and use the adjusted multiple coefficient of determination (R^2_{adj}) and ANOVA for our model evaluation and comparison respectively.

The results show that a logarithmic transformation of the response is needed with non-linear covariates and the interaction terms are equally required.

The results equally show that rent price per square meter increases exponentially over time (2015 and 2019) in both Munich and Berlin. Considering apartments with a kitchen, rent increases in Berlin and Munich by 24% and 31% respectively from 2015 to 2019. However, overall Munich has higher rent prices than Berlin.

The results also show that **rent_sqm** increases when renting larger apartments in Berlin but in Munich, it decreases when renting larger apartments. However, when a pet is allowed, the **rent_sqm** decreases in both cities, while apartments with Upscale furnishing increase rent_sqm in both cities.

Considering the prediction with our four scenarios: scenario 1, scenario 2, scenario 3 and scenario 4 (smaller, small, large, and larger apartments), the **rent_sqm** increased from 2015 to 2019 in Berlin by 8.70%, 38.16%, 29.16%, and 69.47% for smaller, small, large, and larger apartments respectively but in Munich it increased in the same period of time by 1.84%, 9.27%, 70.71%, and 60.13% respectively. This is in agreement with the research work carried out by Thomschke (2015), that rent price increased throughout the entire distribution, but more at higher quantiles.

Also, in Berlin 2015, the **predicted rent_sqm increased** from scenario 1 to scenario 2, scenario 1 to scenario 3, and scenario 1 to scenario 4 by 36.31%, 54.35% and 67.73% respectively while in Munich 2015, it **decreased** by 8.11%, 10.69% and 18.66% respectively. On the other side, in Berlin 2019, the **predicted rent_sqm increased** from scenario 1 to scenario 2, scenario 1 to scenario 3, and scenario 1 to scenario 4 by 73.24%, 83.40% and 317.38% respectively while in Munich 2019, it **decreased** by 1.40%, increased by 49.70% and 27.89% respectively.

Also, **rent_sqm** price relatively increases in Berlin with respect to the order of the cate-

gories of the following features; the number of bedrooms (0-1, 2,>2), the inclusion of warm water consumption in the energy consumption value calculation (No, Yes) and the **energy efficiency categories (Low, Medium, and High)** (vice versa). On the other hand, in Munich, the **rent_sqm** relatively decreases with respect to this order (vice versa).

This thesis is organized as follows: Chapter 2 introduces the mathematical and statistical background for this thesis such as model formulations and assumptions. Chapter 3 provides the data description and management. In Chapter 4, we carry out the exploratory data analysis (EDA). In Chapter 5, we fit different models and choose the optimal one for our model prediction according to the goodness of fit test. In Chapter 6, we discuss the summary of findings, recommendations, and conclusion.

Chapter 2

Multiple linear regression model

In this chapter, we talk about the required mathematical and statistical background for this thesis. We look at linear models, their formulations, assumptions, estimations, validation, predictions and hypothesis testing. More details can be found in Czado and Brechmann (2021).

2.1 Concept of a multiple linear regression model

On many occasions, a relationship is found or suspected to exist between two (or more) variables. Sometimes, one might be interested to investigate if there is a relationship or trend between two or more variables, and if they are, how they are related. In regression, we want to model the relationship between the variable of interest (dependent or response variable), and other given variables (covariates or independent variables), see Fahrmeir et al. (2013). For instance, we may want to know whether there exists a relationship between the number of hours students read in a day (independent variable or covariate) and their performances in the examination (dependent or response variable). The goal of regression analysis is to determine the parameters of the linear function that best describes the joint distribution of the response variable and the covariates (Allen, 2004). We note that the relationship among variables may be linear, nonlinear (quadratic, cubic, etc.) or non-relationship at all and may involve several independent variables. Thus, we need tools for an exploratory data analysis (EDA), which enables us to suggest useful model formulations before fitting specific regression models. We refer to multiple linear regression when several independent variables are involved and the response variable is a continuous variable.

In this study, we want to investigate the relationship between rent per square meter charged for an apartment characterized by a set of continuous and discrete covariates. We will investigate whether we need a transformation of the response and whether covariates enter the model and whether interaction terms are required.

2.2 Model formulation

In a regression analysis with a continuous response variable Y_i and p covariates or predictors $X_{i1}, X_{i2}, \ldots, X_{ik}$ which may be continuous or qualitative (ordinal or nominal) with n observations, let $\left(y_i, \boldsymbol{x}_i^\top\right) := (y_i, x_{i1}, \ldots, x_{ik})^\top, i = 1, \ldots, n$, $k = p - 1$, be a pair of the ith observation $\left(y_i, \boldsymbol{x}_i^\top\right)$ of the random vector $\left(Y_i, \boldsymbol{x}_i^\top\right)$, where $\boldsymbol{x}_i = (x_{i1}, x_{i2}, \ldots, x_{ik})^\top$, then our objective is to analyze the effects of the covariates on the mean value of the response variable ($\mu_i \equiv E[Y_i]$).

The linear model models the response as a linear function of the predictors together plus an error term, i,e.

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \epsilon_i = \beta_0 + \sum_{j=1}^{k} \beta_j x_j + \epsilon_i \tag{2.1}$$

with mean $E[Y_i] = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik}$.

Definition 2.1 *Multiple linear regression model The multiple linear regression model is defined as*

$$Y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik} + \varepsilon_i, \quad i = 1, \ldots, n, \tag{2.2}$$

where ϵ_i is the random error variable, β_0 is the intercept and the k parameters $\beta_1, \ldots, \beta_k$ are the unknown regression parameters which have to be estimated from n observations $(y_i, x_{i1}, \ldots, x_{ik}), i = 1, \ldots, n$.

Matrix representation

It is very easy to represent our linear model in a matrix form (Brown, 2014). The four different model components below, are defined in order to represent the multiple linear regression model of (2.2) in the matrix-vector notation for our model formulation and calculation.

(a) Let $\mathbf{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix} \in \mathbb{R}^n$, be the vector of the response variables.

(b) Let $X = \begin{pmatrix} 1 & x_{11} & x_{12} & \ldots & x_{1k} \\ 1 & x_{21} & x_{22} & \ldots & x_{2k} \\ & & \vdots & & \\ 1 & x_{n1} & x_{n2} & \ldots & x_{nk} \end{pmatrix} \in \mathbb{R}^{n \times p}$, be the design matrix that contains

$p = k + 1$ predictors with their n observations in its rows. The columns correspond to the p unknown regression parameters. Note that the first column which corresponds to the intercept β_0 equals 1 for all n entries.
Denote by $\boldsymbol{x}_i \in \mathbb{R}^p$ the ith row of the design matrix.

(c) Let $\boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix} \in \mathbb{R}^p$, be the unknown vector of the regression coefficients.

(d) Let $\boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix} \in \mathbb{R}^n$, be the vector of random error variables.

Definition 2.2 (*Linear regression model in matrix − vector notation*) *With these notations model (2.2) can be expressed as*

$$\underbrace{\begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}}_{n\times 1} = \underbrace{\begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{pmatrix}}_{n\times p} \underbrace{\begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix}}_{p\times 1} + \underbrace{\begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}}_{n\times 1},$$

Thus, the multiple linear regression (2.2) can now be written as

$$\boldsymbol{Y} = X\boldsymbol{\beta} + \epsilon_i, \text{ with } \epsilon_i \sim N_n\left(\mathbf{0}, \sigma^2 I_n\right) \tag{2.3}$$

Where I_n is the $n \times n$ identity matrix and $N_m(\boldsymbol{\mu}, \Sigma)$ denotes the m-dimensional multivariate normal distribution with the mean vector $\boldsymbol{\mu}$ and covariance matrix Σ.

2.3 Assumptions of the linear model

(a) **Linearity in the covariates**: In (2.1), we introduced that the relationship between the covariate vector $\boldsymbol{x}_i$ and the random response Y_i has the form

$$Y_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} + \varepsilon_i$$

with random error variable ε_i satisfying $E[\varepsilon_i] = 0$, so that

$$E[Y_i] = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}$$

$$i = 1, \ldots, n.$$

In matrix notation: $E[\mathbf{Y}] = \mathbf{X}\boldsymbol{\beta}$.

(b) **Homoscedasticity**: The error variables ε_i have constant variance

$$\mathrm{Var}[Y_i] = \mathrm{Var}[\varepsilon_i] = \sigma^2, \quad i = 1, \ldots, n.$$

(c) **Independence of the random errors**: We assume that the error variables ε_i are independent and identically distributed (i.i.d.). Then it follows

$$\mathrm{Cov}(Y_j, Y_{j'}) = \mathrm{Cov}(\varepsilon_j, \varepsilon_{j'}) = 0, \forall j \neq j'$$

(d) **Normality**: The random error variables ε_i follow a normal distribution.

$$\mathbf{Y} \sim N_n\left(\mathbf{X}\boldsymbol{\beta}, \sigma^2 \mathbf{I}_n\right) \tag{2.4}$$

where I_n is the $n \times n$ identity matrix and $N_m(\boldsymbol{\mu}, \Sigma)$ denotes the m-dimensional multivariate normal distribution with the mean vector $\boldsymbol{\mu}$ and covariance matrix Σ, respectively.

In general, we assume a Gaussian error $\boldsymbol{\varepsilon_i} \sim N_n(0, \sigma^2 \mathbf{I}_n)$. This allows us to construct confidence intervals and conduct statistical tests.

Here, $\boldsymbol{I}_n$ is the n-dimensional identity matrix and $N_m(\boldsymbol{\mu}, \Sigma)$ denotes the m-dimensional multivariate normal distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix Σ, respectively.

2.4 Polynomial regression

Polynomial regression is often appropriate when there exist a relationship between the response and the covariates.

Theorem 2.1 (*Polynomial regression*). *Given a continuous covariate V_i with observations v_i that has a polynomial effect of degree d on the response, then the model $Y_i = \beta_0 + \beta_1 V_i + \beta_2 V_i^2 + \cdots + \beta_d V_i^d + \cdots + \varepsilon_i$ can be used. Note, it is a linear regression model of the form (2.2) with $x_{ij} = v_i^j, j = 1 \ldots d$ (Christensen, 1996) and (Christensen, 2018).*

In order to increase numerical stability, we orthonomalize the corresponding design matrix $X = \begin{pmatrix} 1 & v_1 & v_1^d \\ \vdots & & \\ 1 & v_n & v_n^d \end{pmatrix}$ to X^*, where all columns have unit norms and are orthogonal. In *R*, this is achieved by poly(v, d), see Horton and Kleinman (2015).

2.5 Transformations of the response variable

Sometimes, the transformation of the response variable is appropriate when non-normality and/or unequal error variances are present in the data. We will consider three different transformations of the response variable in this thesis.

Definition 2.3 (*Logarithmic and inverse transformation*) *Given a response variable* $\mathbf{Y}$ *that has an exponential relationship with the covariates. Let* $Y_i^{ln} := \ln(Y_i)$*, let* $Y_i^{lnln} := \ln(\ln(Y_i))$*, let* $Y_i^{inv} := \frac{1}{Y_i}$*, then the formulated model* $Y_i = \exp(\beta_0 + \beta_1 x_{i1}, \ldots, \beta_k x_{ik} + \varepsilon_i)$ *can be expressed in the form of the linear regression model (2.2) as*

$$Y_i^{ln} = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} + \varepsilon_i, \quad i = 1, \ldots, n \tag{2.5}$$

$$Y_i^{lnln} = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} + \varepsilon_i, \quad i = 1, \ldots, n \tag{2.6}$$

$$Y_i^{inv} = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} + \varepsilon_i, \quad i = 1, \ldots, n \tag{2.7}$$

2.6 Interaction effects among the covariates

Interaction effects among the covariates is a situation where two or more covariates have a joint effect on the response variable.

Example 2.1 *A regression model with* $\mathbf{Y}$ *as the response vector,* $\boldsymbol{x}_1 = (x_{i1})_{i=1,\ldots,n}$ *and* $\boldsymbol{x}_2 = (x_{i2})_{i=1,\ldots,n}$ *as the covariates allows for an interaction effect on* $\mathbf{Y}$*, when we consider the following model*

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i1} x_{i2} + \varepsilon_i, \quad i = 1, \ldots, n \tag{2.8}$$

The terms $\beta_1 x_{i1}$ and $\beta_2 x_{i2}$ depend only on $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ respectively, thus are called the **main effects**. On the other hand, the term $\beta_3 x_{i1} x_{i2}$ jointly depends on both $\beta_1 x_{i1}$ and $\beta_2 x_{i2}$, thus is called the **interaction effect** between x_{i1} and x_{i2} on $\mathbf{Y}$.

We can give a clearer interpretation of the interaction term by examining the change of $E[\mathbf{Y}]$ when a variable changes by v units. Thus, by adding v to the first covariate $\boldsymbol{x}_1$, then we get

$$\begin{aligned} E\left[Y_i \mid x_{i1}+v, x_{i2}\right] - E\left[Y_i \mid x_{i1}, x_{i2}\right] =& \beta_0 + \beta_1\left(x_{i1}+v\right) + \beta_2 x_{i2} + \beta_3\left(x_{i1}+v\right)x_{i2} \\ & - \beta_0 - \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i1} x_{i2} \\ =& \beta_1 v + \beta_3 v x_{i2}. \end{aligned} \tag{2.9}$$

We now have two different cases, $\beta_3 = 0$ and $\beta_3 \neq 0$

- $\beta_3 = 0$: The interaction is excluded from the model, but main effects are included. The expected change $\beta_1 v$ is not dependent from the value of the second covariate x_2.
- $\beta_3 \neq 0$: The expected change $\beta_1 v + \beta_3 v x_{i2}$ is dependent on the additional amount v and also on the value of the second predictor $\boldsymbol{x}_2$.

Thus, if the effect of changing a covariate depends on the value of another covariate, then it is necessary to add an interaction term.

Note that we should always test for the presence of an interaction term of the two cases in (2.9) first. If $H_0 : \beta_3 = 0$, cannot be rejected, we can test the significance of the main effects $H_0 : \beta_1 = 0$ or $H_0 : \beta_2 = 0$. If $H_0 : \beta_3 = 0$ is rejected, then both main effects $\beta_1 x_{i1}$ and $\beta_2 x_{i2}$ need to be in the model specification as well as $\beta_3 x_{i1} x_{i2}$.

2.7 Estimation of model parameters

In this section, we will consider the methods of estimating the unknown parameters in the linear regression model of Definition (2.2). Our goal is to determine estimates

$$\widehat{\boldsymbol{\beta}} = \left(\widehat{\beta}_0, \ldots, \widehat{\beta}_k\right)^\top \in \mathbb{R}^p \tag{2.10}$$

and the error variance σ based on n observations. Here $\boldsymbol{\beta}$ is the unknown regression parameter vector.

Note that parameter **estimators**, which are random quantities are different from their realizations called **estimates**, which are determined by the values of the observations. We will consider two approaches, least squares (LS) estimation, and maximum likelihood (ML) estimation. These two estimation methods yield the same estimator if the assumptions of independence, homoscedasticity, and normality of errors are satisfied.

Least squares estimation method

Let the fitted values of the Model (2.2) be given as

$$\begin{aligned} \hat{Y}_i &= \widehat{\beta}_0 + \widehat{\beta}_1 x_{i1} + \ldots + \widehat{\beta}_k x_{ik}, \quad i = 1, \ldots, n \\ &= \boldsymbol{x}_i' \hat{\boldsymbol{\beta}} \end{aligned} \tag{2.11}$$

Also, let the residual denoted by $\hat{\boldsymbol{\varepsilon}} = (\hat{\varepsilon}_1, \ldots, \hat{\varepsilon}_n)' \in \mathbb{R}^n$, which is the difference between the observed response values y_i and the corresponding fitted values of (2.11), be given as

$$\hat{\boldsymbol{\varepsilon}} = \mathbf{y} - \widehat{\mathbf{y}} = \mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}, \tag{2.12}$$

where $\widehat{\boldsymbol{y}} = (\widehat{y}_1, \ldots, \widehat{y}_n)^\top \in \mathbb{R}^n$ in the vector notation. Then, least squares minimizes the residual sum of squares (the sum of the squared deviations) of Equation (2.12).

Definition 2.4 (*Sum of squared deviations*) *Given the data* $(y_i, x_i), i = 1, 2, \ldots, n$*, the sum of the squared deviations which is used in obtaining the estimates* $\widehat{\boldsymbol{\beta}}$ *of Equation (2.10) for the unknown regression parameters* $\boldsymbol{\beta}$ *is given as*

$$Q_{LS}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \left(y_i - \boldsymbol{x}_i^T \boldsymbol{\beta}\right)^2 = \sum_{i=1}^{n} \hat{\varepsilon}_i^2 = \hat{\boldsymbol{\varepsilon}}^T \hat{\varepsilon} \tag{2.13}$$

In order to minimize $Q_{LS}(\boldsymbol{\beta})$ (2.13), we take the partial derivative of $Q_{LS}(\boldsymbol{\beta})$ with respect to $\boldsymbol{\beta}$ and set the result to zero. Then, it follows

$$\frac{\partial(Q_{LS}(\boldsymbol{\beta}))}{\partial \boldsymbol{\beta}} = \mathbf{0} \quad \Leftrightarrow \quad -2\mathbf{X}^T\mathbf{y} + 2\mathbf{X}^T\mathbf{X}\boldsymbol{\beta} = \mathbf{0} \quad \Leftrightarrow \quad X^\top X\boldsymbol{\beta} = X^\top \boldsymbol{y} \tag{2.14}$$

We are now interested in solving the least squares normal equations given in (2.14). If the matrix X has a full rank p, then $\mathbf{X}^T\mathbf{X}$ will be positive definite and will have a unique solution. Thus, the minimum of $Q_{LS}(\boldsymbol{\beta})$ is attained at

$$\widehat{\boldsymbol{\beta}}_{\boldsymbol{LS}} = \left(X^\top X\right)^{-1} X^\top \boldsymbol{y} \tag{2.15}$$

which is the least squares estimate from the normal equations.

Maximum likelihood estimation method

The method of maximum likelihood estimation is based on specifying the distribution we are sampling from and writing the joint density of our sample, unlike in the least squares method where we do not specify the distribution of the response variable Y_i.

Considering the assumptions of our linear model, we assumed in Equation (2.4) that the random variables Y_i are normally distributed ($\mathbf{Y} \sim N_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}_n)$). Thus, It follows that the likelihood of the vector $(\boldsymbol{\beta}, \boldsymbol{\sigma})$ given the data values $\mathbf{y}$ is

$$L(\boldsymbol{\beta}, \boldsymbol{\sigma} \mid \mathbf{y}) = \frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}} \exp\left(-\frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})\right) \tag{2.16}$$

Therefore, the corresponding log likelihood is given by

$$l(\boldsymbol{\beta}, \boldsymbol{\sigma} \mid \mathbf{y}) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\left(\sigma^2\right) - \frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \tag{2.17}$$

To maximize this log-likelihood (2.17) with respect to $\boldsymbol{\beta}$, we differentiate equation (2.17) with respect to $\boldsymbol{\beta}$ and set it equal to zero (Nelder and Wedderburn, 1972). Thus, we have

$$\frac{\partial(l(\boldsymbol{\beta}, \boldsymbol{\sigma} \mid \mathbf{y}))}{\partial\boldsymbol{\beta}} = \mathbf{0} \quad \Leftrightarrow \quad -\frac{1}{2\sigma^2}(-2\mathbf{X}^T\mathbf{y} + 2\mathbf{X}^T\mathbf{X}\boldsymbol{\beta}) = \mathbf{0} \quad \Leftrightarrow \quad X^\top X\boldsymbol{\beta} = X^\top \boldsymbol{y} \tag{2.18}$$

This shows that $\widehat{\boldsymbol{\beta}}_{\boldsymbol{ML}} = \widehat{\boldsymbol{\beta}}_{\boldsymbol{LS}}$.

Also, differentiating Equation (2.17) with respect to σ^2 and maximizing over σ^2, we have

$$\widehat{\sigma}^2 := \frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2 = \frac{1}{n}\|\hat{\boldsymbol{\varepsilon}}\|^2 \tag{2.19}$$

Distribution of the estimators

Definition 2.5 *($\hat{\mathbf{Y}}$ and $\mathbf{H}$) We define the vector of the fitted random values $\hat{\mathbf{Y}}$ as*

$$\widehat{\boldsymbol{Y}} = X\widehat{\boldsymbol{\beta}} = X\left(X^\top X\right)^{-1}X^\top \boldsymbol{Y} \tag{2.20}$$

Also, we define the hat matrix H which gives the projection of the vector $\mathbf{Y}$ onto the space that is spanned by the columns of the design matrix $\mathbf{X}$ as

$$H := X\left(X^\top X\right)^{-1}X^\top \in \mathbb{R}^{n\times n} \tag{2.21}$$

It can be easily be shown in lemma (2.7.3) that H is both symmetric ($H^T = H$) and idempotent ($H^2 = H$) using the fact that $(X'X)^{-1}$ is symmetric ($\left[(X'X)^{-1}\right]' = (X'X)^{-1}$).

Lemma 2.1 *(**Symmetry** and **Idempotence** of* $\mathbf{H}$*) Using $\left[(X'X)^{-1}\right]' = (X'X)^{-1}$, $H = H^2$ and $H' = H$ holds for $H = X\left(X'X\right)X'$.*

Proof 2.1

$H' = \left(X\left(X'X\right)^{-1}X'\right)' = X\left[(X'X)^{-1}\right]'X' = X\left(X'X\right)^{-1}X' = H$, *using* $(AB)' = B'A'$.

$HH = X\left(X'X\right)^{-1}X'X\left(X'X\right)^{-1}X' = H$.

Since the estimators $\hat{\boldsymbol{\beta}}$ of the regression coefficients $\boldsymbol{\beta}$, the fitted values $\hat{\mathbf{Y}}$ and the raw residuals $\hat{\boldsymbol{\varepsilon}}$ are all linear functions of the vector of random variables Y_i, we can apply the transformation rules for expectation and variance-covariance matrix respectively, to show that

$$\begin{array}{ll} E[\hat{\boldsymbol{\beta}}] = \boldsymbol{\beta}, & \mathrm{Var}[\hat{\boldsymbol{\beta}}] = \sigma^2\left(X^\top X\right)^{-1}, \\ E[\hat{\boldsymbol{Y}}] = \mathbf{X}\boldsymbol{\beta}, & \mathrm{Var}[\hat{\boldsymbol{Y}}] = \sigma^2 H, \\ E[\hat{\boldsymbol{\varepsilon}}] = \mathbf{0}, & \mathrm{Var}[\hat{\varepsilon}] = \sigma^2\left(I_n - H\right) \end{array} \tag{2.22}$$

Considering the normality assumption since $\hat{\boldsymbol{\beta}}$, $\hat{\mathbf{Y}}$, and $\hat{\boldsymbol{\varepsilon}}$ are linear functions of $\mathbf{Y}$, we have

$$\begin{array}{l} \hat{\boldsymbol{\beta}} \sim N_p\left(\boldsymbol{\beta}, \sigma^2\left(X^\top X\right)^{-1}\right) \\ \hat{\boldsymbol{Y}} \sim N_n\left(\mathbf{X}\boldsymbol{\beta}, \sigma^2 H\right) \\ \hat{\boldsymbol{\varepsilon}} \sim N_n\left(0, \sigma^2\left(I_n - H\right)\right) \end{array} \tag{2.23}$$

It can be shown that the variance estimator $\hat{\sigma}^2$ given in (2.19) is given by

$$\mathrm{E}\left(\widehat{\sigma}^2\right) = \frac{n-p}{n}\sigma^2.$$

and an unbiased estimator s^2 of σ^2 is given by

$$s^2 := \frac{1}{n-p}\sum_{i=1}^{n}\left(Y_i - \widehat{Y}_i\right)^2 = \frac{n}{n-p}\widehat{\sigma}^2 = \frac{1}{n-p}\|\hat{\boldsymbol{\varepsilon}}\|^2. \tag{2.24}$$

2.8 Goodness of fit and model selection

It is of great importance to know the goodness of the fitted model after estimating the parameters of the linear regression model of (2.2). Thus, we need suitable measures of the goodness of fit. Therefore, we will introduce one of the appropriate measures of the goodness of fit called the coefficient of determination (R^2) which determines the proportion of variation of the response variable that is explained by the covariates.

Sum of squares

Definition 2.6 ***(Sum of squares)*** *We define the sum of squares* SST ***(total sum of squares)***, SSR ***(regression sum of squares)*** *and* SSE ***(error sum of squares)*** *to quantify the amount of variability explained by the regression model as follows*

$$\begin{aligned} \text{SST} &:= \sum_{i=1}^{n} (y_i - \bar{y})^2 \quad \Leftrightarrow \quad (total\ sum\ of\ squares) \\ \text{SSR} &:= \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 \quad \Leftrightarrow \quad (regression\ sum\ of squares) \\ \text{SSE} &:= \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \quad \Leftrightarrow \quad (error\ sum\ of\ squares) \end{aligned} \tag{2.25}$$

where $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$. Thus, we can have the decomposition as

$$\sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \tag{2.26}$$

and using the fact that $\sum_{i=1}^{n} (\hat{y}_i - \bar{y})(y_i - \hat{y}_i) = 0$, it follows from (2.26) that

$$\text{SST} = \text{SSR} + \text{SSE} \tag{2.27}$$

Selection of model (R^2 and adjusted R^2)

The multiple coefficient of determination R^2 is a measure of goodness of fit. It measures how well the covariates in the model explain the variance in the response variable, see Abraham and Ledolter (2006).

Definition 2.7 *(Multiple coefficient of determination) We define the* ***multiple coefficient of determination*** $\mathbf{R}^2$ *as*

$$\text{R}^2 := \frac{\text{SSR}}{\text{SST}} = 1 - \frac{\text{SSE}}{\text{SST}} \tag{2.28}$$

We also define the ***adjusted multiple coefficient of determination*** R^2_{adj} *as*

$$\mathrm{R}^2_{adj} := 1 - \frac{\mathrm{SSE}/(n-p)}{\mathrm{SST}/(n-1)} \tag{2.29}$$

The values of the multiple coefficient of determination range from zero to one ($0 \leq R^2 \leq 1$). Our model accounts for a larger amount of variation of the response when the R^2 is closer to 1. However, the weakness of R^2 is that, it always increases when we add more covariates to our model, and therefore cannot be used to compare the goodness of fit for models with different numbers of covariates, see Ricci (2010). Thus, the need to establish an appropriate measure R^2_{adj} which compares models with different numbers of covariates. We will therefore make use of the adjusted multiple coefficient of determination (R^2_{adj}) as a measure of our model selection in this thesis.

2.9 Hypothesis testing

A statistical hypothesis is an assumption about the form of a population, which based on sample information from the population, seeks to support or reject this assumption. If there is evidence that the null hypothesis (hypothesis of no difference) denoted by H_0 is not true, then it is rejected and its alternative denoted by H_1 is accepted. Thus, a test of hypothesis is a rule or a procedure used for deciding whether to accept or reject H_0 or to determine whether the observed sample differs significantly from expected results under H_0 (McNeil et al., 1996).

This concept can be extended in statistical inference for the model parameters of linear regression (Seber, 2015). For instance, we may want to know if the response variable is significantly influenced by a particular set of covariate variables, which can be expressed in terms of linear combinations of the unknown regression parameters $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_k)^\top$. We will use the Chi-square, F and the univariate t- distribution since the t-test and the F-test rely on quantities of these distributions.

Definition 2.8 *(**Chi-square distribution**) A continuous random variable X is said to have a **Chi-square distribution** with parameter, ν, if its probability density function is given by*

$$f_X(x \mid \nu) = \frac{2^{-\nu/2}}{\Gamma(\nu/2)} x^{\nu/2-1} e^{-x/2}, \nu > 0, x > 0$$

Here, ν is the degree of freedom, $\mathrm{E}(X) = \nu, \mathrm{Var}(X) = 2\nu$. Thus, we say that X follows a Chi-square distribution with ν degree of freedom ($X \sim \chi^2_\nu$).

Definition 2.9 *(**F-distribution**) A continuous random variable X is said to have an **F-distribution** with degrees of freedom (df) ν_1 and ν_2, if its pdf is given by*

$$f(x) = \frac{\Gamma\left(\frac{\nu_1+\nu_2}{2}\right)\left(\frac{\nu_1}{\nu_2}\right)^{\frac{\nu_1}{2}} x^{\frac{\nu_1}{2}-1}}{\Gamma\left(\frac{\nu_1}{2}\right)\Gamma\left(\frac{\nu_2}{2}\right)\left(1+\frac{\nu_1 x}{\nu_2}\right)^{\frac{\nu_1+\nu_2}{2}}}, \quad x \geq 0. \tag{2.30}$$

If $X_1 \sim \chi^2_{v_1}$ and $X_2 \sim \chi^2_{\nu_2}$ and are independent, it follows in (2.30) that X is F-distributed with ν_1 and ν_1 df.

$$X = \frac{X_1/\nu_1}{X_2/\nu_2} \sim F_{\nu_1,\nu_2} \tag{2.31}$$

Definition 2.10 *(**Univariate t-distribution**) A continuous random variable X is said to have a **Univariate t-distribution** with degree of freedom df ν, if its pdf is given by*

$$f_\nu\left(x;\mu,\sigma^2\right) := \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)\sqrt{(\pi\nu)}\sigma}\left\{1+\left(\frac{x-\mu}{\sigma}\right)^2\frac{1}{\nu}\right\}^{-\frac{\nu+1}{2}}, \quad \nu \geq 1 \tag{2.32}$$

$$\mathrm{E}(X) = \mu \text{ and } \mathrm{Var}(X) = \frac{\nu}{\nu-2}\sigma^2.$$

If $X_1 \sim N(0,1)$ and $X_2 \sim \chi^2_n$ and are independent, it can be be shown in (2.32) that T has a t-distribution with ν df.

$$T = \frac{X_1}{\sqrt{\frac{X_2}{\nu}}} \sim t_\nu. \tag{2.33}$$

F-test

Definition 2.11 *(**General testing problem**) We define the general testing problem as*

$$H_0 : C\boldsymbol{\beta} = \boldsymbol{d} \text{ versus } H_1 : C\boldsymbol{\beta} \neq \boldsymbol{d} \tag{2.34}$$

*where matrix $C \in \mathbb{R}^{q\times p}$ with $\mathrm{rank}(C) = q$ and $\boldsymbol{d} \in \mathbb{R}^q$ is called the **general linear hypothesis**.*

Using the distribution of the estimated regression coefficients $\hat{\boldsymbol{\beta}}$ given in (2.23), if H_0 is true, it follows that

$$\widehat{\boldsymbol{\vartheta}} = C\widehat{\boldsymbol{\beta}} - \boldsymbol{d} \overset{H_0}{\sim} N_q\left(\mathbf{0}, \sigma^2 C\left(X^\top X\right)^{-1} C^\top\right). \tag{2.35}$$

We used the fact that

$$\begin{aligned} \mathrm{E}(\widehat{\boldsymbol{\vartheta}}) &= C\mathrm{E}(\widehat{\boldsymbol{\beta}}) - \boldsymbol{d} = C\mathrm{E}(\boldsymbol{\beta}) - \boldsymbol{d} = \mathbf{0} \\ \mathrm{Var}(\widehat{\boldsymbol{\vartheta}}) &= C\,\mathrm{Var}\,\hat{\boldsymbol{\beta}}C^\top = \sigma^2 C\left(X^\top X\right)^{-1} C^\top. \end{aligned} \tag{2.36}$$

Also, using the spectral decomposition for a specific covariance and considering the definition of χ^2-distribution, it can be shown that

$$\frac{1}{\sigma^2}\widehat{\boldsymbol{\vartheta}}^\top \left(C\left(X^\top X\right)^{-1} C^\top\right)^{-1} \widehat{\boldsymbol{\vartheta}} \overset{H_0}{\sim} \chi^2_q \tag{2.37}$$

One can also show that $\frac{1}{\sigma^2}\widehat{\boldsymbol{\vartheta}}^\top \left(C\left(X^\top X\right)^{-1} C^\top\right)^{-1} \widehat{\boldsymbol{\vartheta}} \sim \chi^2_q$ and $\mathrm{SSE}/\sigma^2 \sim \chi^2_{n-p}$, and are independent χ^2 distributed. We therefore define the statistic under H_0 as

$$F = \frac{\widehat{\boldsymbol{\vartheta}}^\top \left(C\left(X^\top X\right)^{-1} C^\top\right)^{-1} \widehat{\boldsymbol{\vartheta}}/q}{\mathrm{SSE}\,/(n-p)} \overset{H_0}{\sim} F_{q,n-p}. \tag{2.38}$$

If the null hypothesis $H_0 : C\boldsymbol{\beta} = \boldsymbol{d}$ holds (ie, $C\boldsymbol{\beta} - \boldsymbol{d} = 0$), then we will reject H_0 for large values of F *since small value of $C\hat{\boldsymbol{\beta}} - \boldsymbol{d}$ is expected.*

Let the least square estimate among those $\boldsymbol{\beta}$ vectors which satisfy $C\boldsymbol{\beta} = \boldsymbol{d}$ be denoted as $\boldsymbol{\beta}_{H_0}$, i.e. it minimizes (2.13) under the condition $\boldsymbol{C\beta} = \boldsymbol{d}$. We define the correponding sum of squares error for the LS fit under H_0 as

$$\mathrm{SSE}_{H_0} := \left\|\boldsymbol{y} - X\widehat{\boldsymbol{\beta}}_{H_0}\right\|^2$$

It can also be shown that $\widehat{\boldsymbol{\vartheta}}^\top \left(C\left(X^\top X\right)^{-1} C^\top\right)^{-1} \widehat{\boldsymbol{\vartheta}} = \mathrm{SSE}_{H_0} - \mathrm{SSE}$ is true which allows us to give the general F-test.

Definition 2.12 ***(General F test in linear regression)*** *We define the test statistic $\boldsymbol{F}$ under the null hypothesis $H_0 : C\boldsymbol{\beta} = \boldsymbol{d}$ versus $H_1 : C\boldsymbol{\beta} \neq \boldsymbol{d}$ in the linear regression model of* (2.2) *as*

$$F = \frac{\left(\mathrm{SSE}_{H_0} - \mathrm{SSE}\right)/q}{\mathrm{SSE}/(n-p)} \overset{H_0}{\sim} F_{q,n-p} \tag{2.39}$$

We reject H_0 against H_1 at level α if

$$F > F_{(1-\alpha),q,n-p} \tag{2.40}$$

Here, $F_{(1-\alpha),q,n-p}$ is the $(1-\alpha)$ quantile of an F *distribution with q and $n-p$ df. The quantity $n-p$ is also called the* ***residual degree of freedom****. Thus, we can now summarize the F-test procedure for our model (2.2) as follows*

Hypothesis

$$H_0 : (\beta_1, \ldots, \beta_k) = \mathbf{0}$$
$$H_1 : (\beta_1, \ldots, \beta_k) \neq \mathbf{0}$$

Test statistic *= F, defined in (2.40)*

Rejection Rule*: Reject H_0 at level α, if $F > F_{(1-\alpha),q,n-p}$*

t-test

Definition 2.13 ***(t-test)*** *We define the* ***t-test procedure*** *for our model (2.2) as follows, since in a t-test, the test statistic is computed for each β_j, see Vik (2013).*

Hypotheses :

$$H_0 : \beta_j = 0 \text{ versus } H_1 : \beta_j \neq 0$$

Test statistic :

$$T_j = \frac{\widehat{\beta}_j}{\widehat{\text{se}}\left(\widehat{\beta}_j\right)} \sim t_{n-p}, \text{under } H_0 \tag{2.41}$$

Here, $\widehat{\text{se}}\left(\widehat{\beta}_j\right) := s\sqrt{\left((X^\top X)^{-1}\right)_{jj}}$ is the estimated standard error of $\widehat{\beta}_j$ and $s = \sqrt{s^2}$ defined in Equation (2.24)

Rejection Rule*: Reject H_0 at level α, if $|T_j| > t_{n-p,1-\alpha/2}$*

Note that **t-test** is a special case of the **F-test**, in particular we have $F_j = T_j^2$ since if

$$\begin{aligned} F_j &:= \frac{\widehat{\beta}_j^2}{\left((X^\top X)^{-1}\right)_{jj} \text{SSE}/(n-p)} \overset{H_0}{\sim} F_{1,n-p} \\ T_j &:= \frac{\widehat{\beta}_j}{\widehat{se}\left(\widehat{\beta}_j\right)} \overset{H_0}{\sim} t_{n-p} \end{aligned} \tag{2.42}$$

Analysis of Variance (ANOVA)

Definition 2.14 ***ANOVA*** *is mostly used to summarize the hypothesis tests results in linear models in a tabular form. Given two models* M_{reduced} *and* M_{full} *which are nested:* $M_{\text{reduced}} \subset M_{\text{full}}$*, that is, all covariates of the reduced model are contained in the full model, we define the* ***ANOVA-test ratio*** *for the comparison of* M_{reduced} *and* M_{full} *as follows*

$$F = \frac{(SSE_{reduced} - SSE_{full}) / (n - p_{full})}{SSE_{reduced} / (p_{full} - p_{reduced})} \sim F_{n-p_{full}, p_{full}-p_{reduced}} \tag{2.43}$$

Now we compare two models, one with only main effects denoted here as ***reduced*** *and the other one with interaction terms included denoted as* ***full****. For this, we run an* ***ANOVA-ratio test*** *to test if the model with the interaction terms is significantly better than the model with only the main effect. Here, we defined* β_3 *in Equation (2.9) as the interaction effect for the full model, thus we summarize the* ***ANOVA-ratio test procedure*** *as follows*

Hypotheses

$$H_0 : \beta_3 = 0 \text{ versus } H_1 : \beta_3 \neq 0$$

Test statistic : *F, defined in Equation (2.43)*

Rejection Rule*: Reject* H_0 *at level* α*, if* $F > F_{(1-\alpha),n-p_{full},p_{full}-p_{reduced}}$

2.10 Analysis of residuals

After estimating the model parameters, the credibility of the assumptions of linearity, normality of errors, and homoscedasticity for the given data can be assessed using residuals. It is therefore of importance to study the residual in order to examine in what extent our model assumptions may be violated. Therefore, taking a look at the patterns in the residual plots could help us understand if our model assumptions are violated or not. This is called analysis of residual. Residual plots can equally help us to decide whether to transform any of the covariates which we may want to include in the model or not. We will introduce three types of residuals for the *ith* observation namely: **raw residuals**, **internally studentized residuals** and **externally studentized residuals**.

Raw residuals check

The **raw residual vector** $\hat{\boldsymbol{\varepsilon}}$ was defined in (2.12), and its distribution in (2.22). Thus, we have that

$$\hat{\boldsymbol{\varepsilon}} = \mathbf{y} - \widehat{\mathbf{y}} = \mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{Y} - \underbrace{X\left(X'X\right)^{-1}X'}_{H}\boldsymbol{Y} = \boldsymbol{Y} - H\boldsymbol{Y} = \left(I_n - H\right)\boldsymbol{Y}, and$$

$$\mathrm{Var}\left(\hat{\boldsymbol{\varepsilon}}_i\right) = \sigma^2\left(1 - h_{ii}\right), \tag{2.44}$$

where $h_{ii} = \boldsymbol{x}_i^\top \left(X^\top X\right)^{-1} \boldsymbol{x}_i$ is the i th diagonal element of the hat-matrix H defined in (2.21).

Since $\mathrm{Var}\left(\hat{\boldsymbol{\varepsilon}}_i\right)$ still changes under the homoscedasticity condition, we need to standardize the raw residuals. This standardization gives rise to both **internally studentized residuals** and **externally studentized residuals**

Internally studentized residuals

The internally studentized residuals are defined by

$$s_i := \frac{\hat{\boldsymbol{\varepsilon}}_i}{\sqrt{1 - h_{ii}}s}, \tag{2.45}$$

here, s^2 is the estimate of σ^2 defined in (2.19). One can now analyze the variances and conclude whether the assumption of homoscedasticity is violated or not using the standardized residuals by plotting the standardized residuals versus the predicted values $\hat{y}_i$.

However, the deficiency of the internally studentized residuals is that it is not robust against outliers since all data is used to estimate σ and Equation (2.45) is t-distributed. This leads to the definition of externally studentized residuals which is more robust against outliers.

Externally studentized residuals or jackknifed residuals

To solve the problem of the non robustness of the internally studentized residuals, we define a new model just like the model of (2.3), but it is based on "drop-one-observation" of the data, which contains all observations except the i_{th} observation as

$$\boldsymbol{Y}_{-i} = X_{-i}\boldsymbol{\beta}_{-i} + \boldsymbol{\varepsilon}_{-i}, \tag{2.46}$$

here, X_{-i} denotes the design matrix without the ith row, and $\boldsymbol{Y}_{-i}$ is the response vector $\boldsymbol{Y}$ with the ith observation y_i removed. We define the fitted values corresponding to the model given in (2.46) as

$$\widehat{y}_{i,-i} := \boldsymbol{x}_i^\top \widehat{\boldsymbol{\beta}}_{-i} \tag{2.47}$$

where $\widehat{\boldsymbol{\beta}}_{-i}$ is the associated least squares estimates of $\boldsymbol{\beta}_{-i}$. We define the corresponding residual also called the ith **predictive residual** as

$$\hat{\boldsymbol{\varepsilon}}_{i,-1} := y_i - \widehat{y}_{i,-i} \tag{2.48}$$

Using (2.48), we obtain the estimate s^2_{-i} of the error variance σ^2 which does not include the ith observation as

$$s^2_{-i} := \frac{\sum_{j=1,j\neq i}^{n} \left(y_j - \boldsymbol{x}_j^\top \widehat{\boldsymbol{\beta}}_{-i}\right)^2}{n-p-1} \tag{2.49}$$

Finally, we now define the externally studentized residuals which is also called jackknifed residual.

The externally studentized residuals t_i which is based on the "drop-one-observation" of the data, are defined as

$$t_i := \frac{r_{i,-i}}{\sqrt{1-h_{ii}}s_{-i}} \tag{2.50}$$

2.11 Statistical checks for the plausibility of the linear model assumptions

Linearity

The check we are going to use is the residuals versus the fitted values plot. if this plot has no trend , then we assume the linearity assumption as plausible (Lin et al., 2002).

Homoscedasticity

We are interested in checking if $\text{Var}[Y_i] = \text{Var}[\varepsilon_i] = \sigma^2$ holds.

To check this, we use again the standardized residual versus the residual plots. If the standardized residuals are not spread equally along the range of the fitted values, then we interpret the homoscedasticity assumption as not plausible, see Osborne and Waters (2002).

Independence

To check if $\text{Cov}\left(\varepsilon_j, \varepsilon_j'\right) = \rho = 0$ holds, we plot the residuals versus the covariates to see if the residuals are randomly and symmetrically distributed around zero. If this is true, we assume that the independence assumption is plausible (Lin et al., 2002).

Normality

To check for $\varepsilon_i \sim N_n(0, \sigma^2 \mathbf{I}_n)$, we use the Quantile versus Quantile plot (QQPlot). If we do not have a straight line on the QQ plots of our variable versus the theoretical normal quantile, then we assume that the normality assumption is not plausible (Lindsey, 2000).

Chapter 3

Data description and management

3.1 Data description

The data was provided by the FDZ Ruhr at RWI (and ImmobilienScout24) institution. The ImmobilienScout24 GmbH, founded in 1998, deals with real estate properties in Germany. It is located at Invalidenstrasse 65, 10557, Berlin, Germany. The data set contains 2,651,885 observations and 59 attributes from 2007 to 2020. The data has both quantitative and qualitative covariates with **rent per square meter (rent_sqm)** as the response variable. We focus on the most relevant 31 variables such as "the additional cost", "heat cost", "construction year", etc,. The quantitative covariates are summarized as follows: Min = Minimum, 25% = 1st quartile, 50% = Median, $\bar{X}$ = Mean, 75% = 3rd quartile, Max = Maximum and Not available = NA. On the other hand, the qualitative covariates are summarized with their respective categories. Note that costs are expressed in EUR and rounded to two decimal digits and the following data summaries in Table 3.1 and Table 3.2 represent the whole data set. Additionally, adat is the month during which a property was first advertised, while edat is the month where the advertisement was ended.

Table 3.1: Description of quantitative variables

Variables	Description
rent_sqm	Calculated rent per sqm by rent and size of apartment. Min = 3, 25% = 7, 50% = 9, $\bar{X}$ = 9.39, 75% = 12, Max = 28
addcost	The extra monthly costs that need to be paid for other bills on top of the base rent excluding electricity. Min = 0, 25% = 100, 50% = 140, $\bar{X}$ = 153.8, 75% = 196, Max = 599, NA = 97186
heatcost	The monthly heating cost. Min = 0, 25% = 50, 50% = 70, $\bar{X}$ = 75.2, 75% = 94, Max = 300, NA = 898984
conyear	The year in which the object was built Min = 1851, 25% = 1930, 50% = 1970, $\bar{X}$ = 1964, 75% = 1996, Max = 2020, NA = 447372
lmod	The year of the last modernization Min=1981, 25% =2009, 50%=2012, $\bar{X}$=2011, 75%=2015, Max=2018, NA=1113056
lspace	Living space in square meters Min = 19, 25% = 53, 50% = 68, $\bar{X}$ = 71.15, 75% = 85, Max = 165
fspace	The usable floor space in square meters Min = 0, 25% = 16, 50% = 57, $\bar{X}$ = 54.8, 75% = 79, Max = 250, NA = 1053922
energycon	The energy consumption per year and square meter in kWh Min = 0, 25% = 82, 50% = 117, $\bar{X}$ = 120.4, 75% = 152, Max = 350, NA = 977343
adlength	The difference between edat and adat. Min = 0, 25% = 0, 50% = 0, $\bar{X}$ = 0.71, 75% = 1, Max = 20

Table 3.2: Description of qualitative variables

Variables	Description
afloor	Apartment-specific variable indicates the floor the apartment is located in. **afloorg** is used to group afloor as follows: (-1)-0, 1-2, 3-9, >9, NA
bfloor	This indicates the number of floors in the building. **bfloorg** is used to group bfloor as follows: 0-2, 3, 4, 5, >5, NA
nrooms	Number of rooms, excluding kitchen, bath or corridors. **nroomsg** is used to group nrooms as follows: 1-1.5, 2-2.5, 3-3.5, >3.5, NA
nbed	Number of bedrooms of the property. **nbedg** is used to group nbed as follows: 0-1, 2, >2, NA
nbath	Number of bathrooms of the property **nbathg** is used to group nbath as follows: 0-1, > 1, NA
elevator	This variable indicates if a property has an elevator. **elevatorg** is used to group elevator as follows: Yes, No, NA
balcony	This variable indicates the presence of a balcony. **balconyg** is used to group balcony as follows: Yes, No, NA
kitchen	This variable indicates the presence of a fitted kitchen. **kitcheng** is used to group kitchen as follows: Yes, No, NA
eww	if the warm water consumption was included in the energy consumption value calculation. **ewwg** variable is used to group eww as follows: Yes, No, NA
subh	It indicates if a certificate of eligibility to public housing is needed to rent the apartment. **subhg** is used to group subh as follows: Yes, No, NA
gtoilet	This indicates the presence of a guest toilet. **gtoiletg** is used to group gtoilet as follows: Yes, No, NA
garden	This indicates the presence of a garden. **gardeng** is used to group garden as follows: Yes, No, NA
hww	if the warm water consumption was included in the heating cost value calculation. **hwwg** is used to group hww as follows: Yes, No, NA
cellar	This indicates if an property has a cellar room **cellarg** is used to group cellar as follows: Yes, No, NA
parking	This variable indicates whether a parking space is available. **parking** is used to group parking as follows:: Yes, No, NA
furnishing	This is an artificial category number indicating the facilities of the property. **furnishingg** is used to group furnishing as follows: (Upscale, Luxury) = Upscale, (Normal, Simple) = Normal, no specification = NA
eeff	This indicates the energy efficiency rating. **eeffg** is used to group eeff as follows: (A, APLUS, B) = High, (C, D, E) = Medium, (F, G, H) = Low, no specification = NA
ecert	The type of energy performance certificate that the customer has for the object **ecertg** is used to group ecert as follows: Final energy demand = building, Energy consumption characteristic = consumption, NA
pets:	This indicates whether pets are allowed in the property. **petsg** is used to group pets as follows: (Yes, by Agreement) = Yes, No = No, no specification = NA
heat	This indicates the type of heating. **heatg** is used to group heat as follows: Central Heating (CH), Non Central Heating (NCH), NA
apcat	This variable categorizes the property into different classes. **apcatg** is used to group apcat as follows: (Penthouse, Maisonette, Attic Apartment) = top, Apartment = middle, (Mezzanine, Terrace apartment) = low, Basement = below, NA
pcon	This indicates the condition of a property. **pcong** is used to group pcon as follows: (First occupancy, First occupancy after renovation) = First, (Maintained, as good as new) = Mt, In need of renovation = Inr, (Modernized, Renovated, Fully Renovated)= Md, NA

3.2 Data sets

We split the date set described in Section 3.1 into four sub data sets: Berlin 2015, Berlin 2019, Munich 2015 and Munich 2019.

The number of rental properties contained in each data set is given in Table 3.3, while the summaries of the response variable is given in Tables 3.4. The summaries of the quantitative covariates is given in Table 3.5 while in Table 3.6 and Table 3.7, we give the summary of each qualitative variable followed by their percentages. The order of each summary table is the following: Berlin 2015, Berlin 2019, Munich 2015 and Munich 2019 respectively.

Table 3.3: Number of rental properties in the four data sets

city	2015	2019
Berlin	49724	49536
Munich	14449	14776

3.3 Univariate data summaries

Table 3.4: Summary of the response variable - rent_sqm for the four data sets

rent_sqm	Min	25%	50%	Mean	75%	Max	NA's
Berlin 2015	3.00	7.00	8.00	8.66	10.00	17.00	0
Berlin 2019	4.00	8.00	11.00	11.62	14.00	27.00	0
Munich 2015	3.00	12.00	13.00	12.91	15.00	17.00	0
Munich 2019	4.00	16.00	18.00	18.32	21.00	28.00	0

Table 3.5: Univariate data summaries of quantitative covariates: first row = Berlin 2015 , second row = Berlin 2019, third row = Munich 2015, fourth row = Munich 2019

Variable	**Summary**						
addcost:	Min	25%	50%	Mean	75%	Max	NA's
	0.00	97.00	140.00	154.24	195.00	592.00	2021
	0.00	100.00	141.00	157.21	200.00	599.00	1070
	0.00	107.00	153.00	164.04	210.00	540.00	1355
	0.00	120.00	170.00	175.47	220.00	550.00	533
heatcost							
	0.00	54.00	75.00	81.18	100.00	300.00	21126
	0.00	49.00	65.00	71.44	90.00	300.00	19077
	0.00	60.00	85.00	89.35	110.00	288.00	10075
	0.00	55.00	80.00	84.84	109.00	300.00	11155
conyear							
	1851	1910	1961	1954	1992	2016	7647
	1853	1918	1972	1964	1998	2020	6437
	1860	1962	1976	1976	1999	2017	3522
	1858	1965	1985	1982	2014	2020	3068
lmod							
	1983	2012	2014	2012	2015	2016	34384
	1982	2013	2016	2014	2018	2018	42152
	1981	2011	2014	2012	2015	2016	9313
	1983	2013	2015	2014	2017	2018	11386
lspace							
	23.00	55.00	69.00	73.80	89.00	161.00	0
	19.00	52.00	65.00	68.64	82.00	158.00	0
	23.00	55.00	71.00	73.79	90.00	161.00	0
	19.00	51.00	67.00	68.54	84.00	157.00	0
fspace							
	0.00	50.00	67.00	69.51	89.00	220.00	35475
	0.00	48.00	65.00	67.69	87.00	250.00	41068
	0.00	10.00	55.00	53.40	81.00	234.00	9276
	0.00	11.00	55.00	53.14	82.00	249.00	11483
energycon							
	0.00	88.00	117.00	121.47	149.00	350.00	16665
	0.00	74.00	105.00	110.69	140.00	347.00	13863
	0.00	85.00	122.00	122.53	155.00	338.00	5975
	0.00	64.00	103.00	104.11	137.00	339.00	7379
adlength							
	0.00	0.00	0.00	0.71	1.00	20.00	0
	0.00	0.00	0.00	0.49	1.00	20.00	0
	0.00	0.00	0.00	0.58	1.00	20.00	0
	0.00	0.00	0.00	0.53	1.00	20.00	0

Table 3.6: Univariate data summaries of qualitative covariates (part I): first row = Berlin 2015 , second row = Berlin 2019, third row = Munich 2015, fourth row = Munich 2019

Variable	Categories					
afloorg	(-1)-0	1-2	3-9	>9	NA	
	4174 (0.08%)	18918 (0.38%)	19719 (0.4%)	868 (0.02%)	6045 (0.12%)	
	4320 (0.09%)	18759 (0.38%)	21016 (0.42%)	1020 (0.02%)	4421 (0.09%)	
	1762 (0.12%)	6732 (0.47%)	3848 (0.27%)	63 (0%)	2044 (0.14%)	
	1648 (0.11%)	6428 (0.44%)	4687 (0.32%)	97 (0.01%)	1916 (0.13%)	
bfloorg	0-2	3	4	5	>5	NA
	3222 (0.06%)	4938 (0.1%)	10313 (0.21%)	8974 (0.18%)	8030 (0.16%)	14247 (0.29%)
	2659 (0.05%)	4426 (0.09%)	8838 (0.18%)	9291 (0.19%)	8698 (0.18%)	15624 (0.32%)
	2744 (0.19%)	2226 (0.15%)	2770 (0.19%)	1833 (0.13%)	1418 (0.1%)	3458 (0.24%)
	2741 (0.19%)	2187 (0.15%)	2517 (0.17%)	2160 (0.15%)	1950 (0.13%)	3221 (0.22%)
nroomsg	1-1.5	2-2.5	3-3.5	>3.5		
	7283 (0.15%)	20391 (0.41%)	15612 (0.31%)	6438 (0.13%)		
	8959 (0.18%)	21660 (0.44%)	14465 (0.29%)	4452 (0.09%)		
	2157 (0.15%)	5710 (0.4%)	4841 (0.34%)	1741 (0.12%)		
	2836 (0.19%)	5949 (0.4%)	4768 (0.32%)	1223 (0.08%)		
nbedg	0-1	2	>2	NA		
	14796 (0.3%)	8465 (0.17%)	3710 (0.07%)	22753 (0.46%)		
	10951 (0.22%)	6107 (0.12%)	2204 (0.04%)	30274 (0.61%)		
	5636 (0.39%)	3562 (0.25%)	1240 (0.09%)	4011 (0.28%)		
	3884 (0.26%)	2329 (0.16%)	684 (0.05%)	7879 (0.53%)		
nbathg:	0-1	>1	NA			
	28941 (0.58%)	3681 (0.07%)	17102 (0.34%)			
	25237 (0.51%)	3151 (0.06%)	21148 (0.43%)			
	10690 (0.74%)	1669 (0.12%)	2090 (0.14%)			
	11310 (0.77%)	1657 (0.11%)	1809 (0.12%)			
elevatorg	Yes	No	NA			
	17145 (0.34%)	30648 (0.62%)	1931 (0.04%)			
	21019 (0.42%)	28517 (0.58%)	0 (0%)			
	6125 (0.42%)	8108 (0.56%)	216 (0.01%)			
	7929 (0.54%)	6847 (0.46%)	0 (0%)			
balconyg:	Yes	No	NA			
	33799 (0.68%)	15207 (0.31)%)	718 (0.01%)			
	35112 (0.71%)	14424 (0.29%)	0 (0%)			
	10863 (0.75%)	3406 (0.24%)	180 (0.01%)			
	11554 (0.78%)	3222 (0.22%)	0 (0%)			
kitcheng:	Yes	No	NA			
	23510 (0.47%)	24111 (0.48%)	2103 (0.04%)			
	23390 (0.47%)	26146 (0.53%)	0 (0%)			
	8756 (0.61%)	5438 (0.38%)	255 (0.02%)			
	9878 (0.67%)	4898 (0.33%)	0 (0%)			
ewwg:	Yes	No	NA			
	12923 (0.26%)	36042 (0.72%)	759 (0.02%)			
	4701 (0.09%)	3449 (0.07%)	41386 (0.84%)			
	3775 (0.26%)	10454 (0.72%)	220 (0.02%)			
	1419 (0.1%)	723 (0.05%)	12634 (0.86%)			
subhg:	Yes	No	NA			
	1123 (0.02%)	43181 (0.87%)	5420 (0.11%)			
	3550 (0.07%)	45986 (0.93%)	0 (0%)			
	30 (0.00%)	12534 (0.87%)	1885 (0.13%)			
	162 (0.01%)	14614 (0.99%)	0 (0%)			
gtoiletg:	Yes	No	NA			
	6404 (0.13%)	43237 (0.87%)	83 (0.00%)			
	4995 (0.10%)	44541 (0.90%)	0 (0%)			
	3186 (0.22%)	11254 (0.78%)	9 (0.00%)			
	2948 (0.20%)	11828 (0.80%)	0 (0%)			

Table 3.7: Univariate data summaries of qualitative covariates (part II): first row = Berlin 2015 , second row = Berlin 2019, third row = Munich 2015, fourth row = Munich 2019

Variable		Categories			
gardeng	Yes	No	NA		
	6740 (0.14%)	39412 (0.79%)	3572 (0.07%)		
	5250 (0.11%)	44286 (0.89%)	0 (0%)		
	2726 (0.19%)	11173 (0.77%)	550 (0.04%)		
	3074 (0.21%)	11702 (0.79%)	0 (0%)		
hwwg:	Yes	No	NA		
	23355 (0.47%)	23785 (0.48%)	2584 (0.05%)		
	24338 (0.49%)	24114 (0.49%)	1084 (0.02%)		
	8856 (0.61%)	4320 (0.3%)	1273 (0.09%)		
	10161 (0.69%)	4088 (0.28%)	527 (0.04%)		
cellarg:	Yes	No	NA		
	27227 (0.55%)	22194 (0.45%)	303 (0.01%)		
	24999 (0.50%)	24537 (0.50%)	0 (0%)		
	11315 (0.78%)	3036 (0.21%)	98 (0.01%)		
	11533 (0.78%)	3243 (0.22%)	0 (0%)		
parkingg:	Yes	No	NA		
	113 (0.00%)	2 (0.00%)	49609 (1.00%)		
	8133 (0.16%)	486 (0.01%)	40917 (0.83%)		
	59 (0.00%)	0 (0%)	14390 (1.00%)		
	7911 (0.54%)	228 (0.02%)	6637 (0.45%)		
furnishingg:	Upscale	Normal	NA		
	15678 (0.32%)	11993 (0.24%)	22053 (0.44%)		
	14417 (0.29%)	8664 (0.17%)	26455 (0.53%)		
	5699 (0.39%)	3591 (0.25%)	5159 (0.36%)		
	7156 (0.48%)	2726 (0.18%)	4894 (0.33%)		
eeffg:	High	Medium	Low	NA	
	796 (0.02%)	1084 (0.02%)	256 (0.01%)	47588 (0.96%)	
	1434 (0.03%)	2039 (0.04%)	348 (0.01%)	45715 (0.92%)	
	314 (0.02%)	257 (0.02%)	63 (0%)	13815 (0.96%)	
	474 (0.03%)	388 (0.03%)	50 (0%)	13864 (0.94%)	
ecertg:	building	consumption	NA		
	11362 (0.23%)	22607 (0.45%)	15755 (0.32%)		
	15767 (0.32%)	20771 (0.42%)	12998 (0.26%)		
	2898 (0.20%)	6027 (0.42%)	5524 (0.38%)		
	3228 (0.22%)	4393 (0.30%)	7155 (0.48%)		
petsg:	Yes	No	NA		
	1603 (0.03%)	3479 (0.07%)	44642 (0.90%)		
	16251 (0.33%)	5250 (0.11%)	28035 (0.57%)		
	947 (0.07%)	4123 (0.29%)	9379 (0.65%)		
	3460 (0.23%)	5629 (0.38%)	5687 (0.38%)		
heatg:	CH	NCH	NA		
	24827 (0.50%)	14874 (0.30%)	10023 (0.20%)		
	17205 (0.35%)	20432 (0.41%)	11899 (0.24%)		
	8056 (0.56%)	3744 (0.26%)	2649 (0.18%)		
	6589 (0.45%)	5560 (0.38%)	2627 (0.18%)		
apcatg:	top	middle	low	below	NA
	4884 (0.10%)	33758 (0.68%)	1444 (0.03%)	188 (0.00%)	9450 (0.19%)
	4372 (0.09%)	34608 (0.70%)	2289 (0.05%)	389 (0.01%)	7878 (0.16%)
	2011 (0.14%)	7627 (0.53%)	515 (0.04%)	80 (0.01%)	4216 (0.29%)
	2066 (0.14%)	7977 (0.54%)	1158 (0.08%)	130 (0.01%)	3445 (0.23%)
pcong:	First	Mt	Md	Inr	NA
	9013 (0.18%)	12680 (0.26%)	12731 (0.26%)	362 (0.01%)	14938 (0.30%)
	9409 (0.19%)	11535 (0.23%)	11289 (0.23%)	364 (0.01%)	16939 (0.34%)
	1781 (0.12%)	5525 (0.38%)	3280 (0.23%)	17 (0%)	3846 (0.27%)
	2682 (0.18%)	5537 (0.37%)	3175 (0.21%)	11 (0%)	3371 (0.23%)

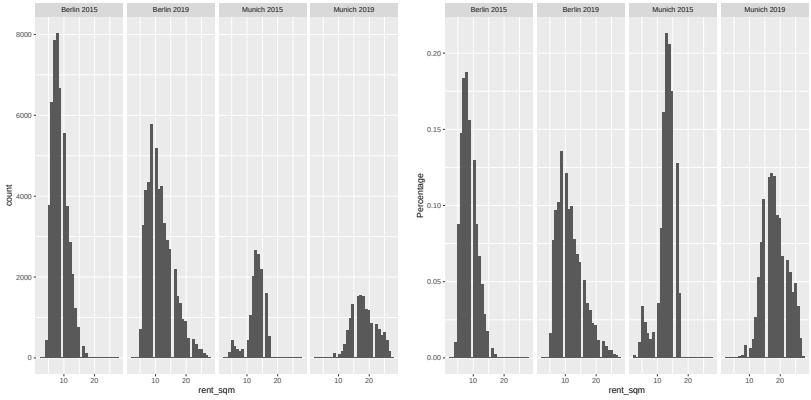
Chapter 4

Exploratory Data Analysis (EDA)

4.1 Histograms of quantitative variables for Berlin and Munich rental properties in 2015 and 2019

The histograms display the raw counts and their respective percentages beside it. Also, the histograms of other continuous variables can be found in the additional EDA plots, Figure A.2 and A.3 of Appendix A

Figure 4.1: Histograms of response variable - **rent_sqm**: first column = **counts**, second column = **percentage**

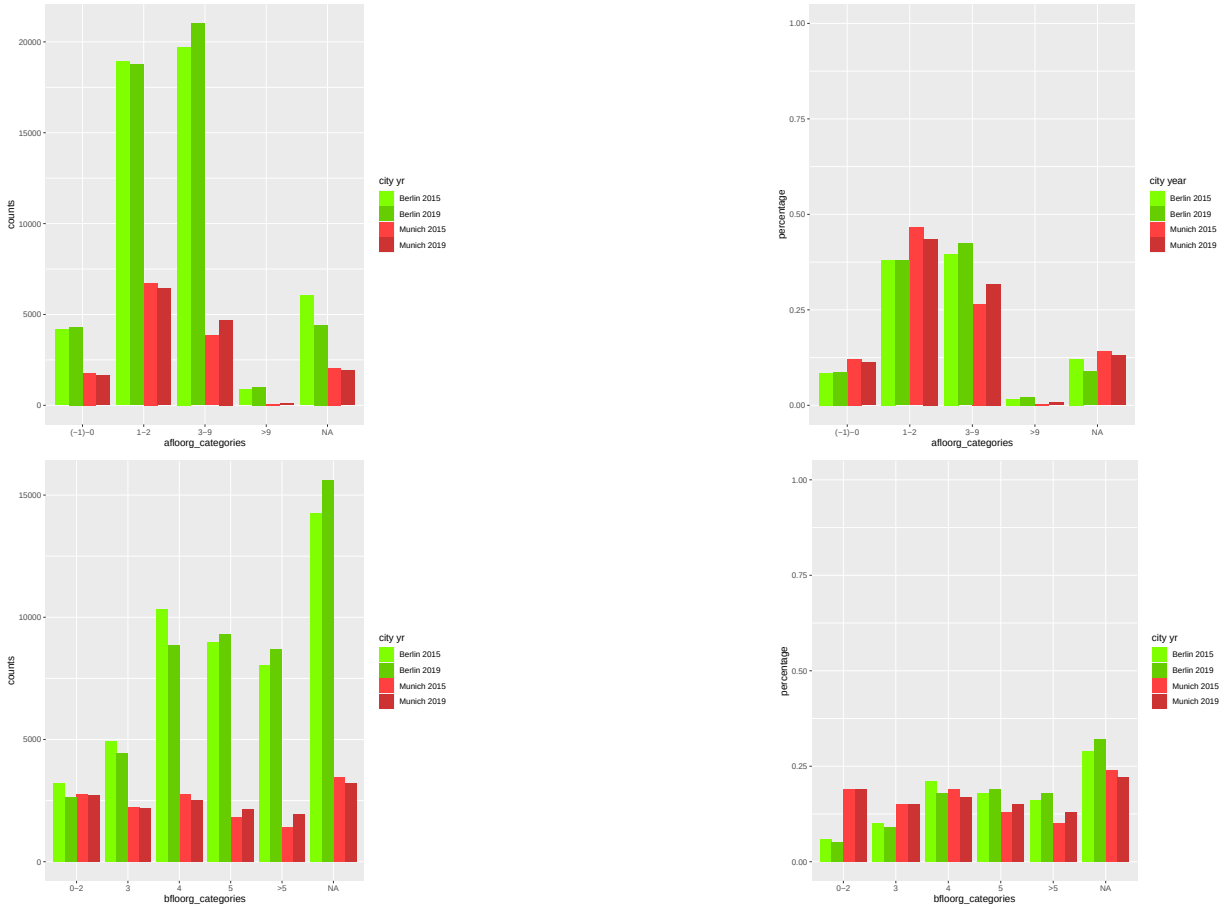


There is a significant shift in the histogram plots of **rent_sqm** for Berlin 2015 and Berlin 2019 as well as in Munich 2015 and Munich 2019. For instance, in Berlin 2015 and Munich 2015, we can see that the **rent_sqm** is below 20 Euros but in 2019, the **rent_sqm** is over 20 Euros for both cities. However, for other continuous variables (addcost, heatcost, etc.), we did not notice a significant shift in the histogram plots for Berlin 2015 and Munich 2015 for the two cities Berlin and Munich 2019.

4.2 Bar plots of qualitative variables for Berlin and Munich rental properties in 2015 and 2019

The bar plots of the qualitative variables for the raw counts and the percentages are plotted on the first and second columns respectively. Also, the bar plots of other qualitative variables can be found in Figure A.4, A.5, A.6, A.7 and A.8 of Appendix A.

Figure 4.2: Bar plot of qualitative variables (part I)



Interpretation of bar plots for Berlin and Munich rental properties in 2015 and 2019

In the above EDA, the results for the qualitative variables are summarized in Table 4.1

Table 4.1: Interpretation of plots for Berlin and Munich rental properties in 2015 and 2019

Variables	Interpretation
afloor	There are more apartments in Munich with **apartment floor (-1-0) and 1-2** compared to Berlin but more apartments in Berlin with **apartment floor 3-9 and >9** compared to Munich
bfloor	In Munich, there are more apartments with **0-3 building floors** compared to Berlin. Conversely, there are more apartments with **5 and >5 building floors** in Berlin compared to Munich
nrooms	Berlin has more apartments with **2-2.5 rooms** compared to Munich but less apartments with **3-3.5 rooms** compared to Munich.
nbed	In Munich, there are more apartments with a higher number of **bedrooms** compared to Berlin
nbath	There are more apartments with a higher number of **bathrooms** in Munich compared to Berlin
elevator	Berlin has more apartments **without an elevator** compared to Munich but Munich 2019 has more apartments **with elevator** compared to other cities.
balcony	Munich has more apartments **with a balcony** compared to Berlin.
kitchen	Munich has more apartments **with a kitchen** compared to Berlin.
eww	There are more apartments **with the inclusion of warm water consumption in the energy consumption value calculation** in both Berlin 2015 and Munich 2015 compared with Berlin 2019 and Munich 2019
subh	There are more apartments in Berlin **with certificate of eligibility to public housing** compared to Munich.
gtoilet	In Berlin, there are more apartments **with guest toilet** and **without guest toilet** compared to Munich.
garden	In Munich, there are more apartments **with garden** and lesser apartments **without garden** compared to Berlin
hww	There are more apartments in Munich **with the inclusion of warm water consumption in the heating cost value calculation** compared to Berlin.
cellar	Munich has more apartments **with a cellar** compared to Berlin:
parking	In Munich, there are more apartments **with a parking space** compared to Berlin
furnishing	Munich has more **upscale furnishing** and **normal furnishing** apartment compared to Berlin
eeff	In Berlin, there are more apartments **with medium energy efficiency rating** compared to Munich.
ecert	Berlin has more apartments with **building type of energy performance certificate** compared to Munich
pets:	There are more apartments in Munich that do **not allow pets** compared to Berlin.
heat	In Munich, more apartments make use of **central heating** as their **heating type** compared to Berlin.
apcat	There are more **middle category** apartments in Berlin compared to Munich but more **top category** apartments in Munich compared to Berlin.
pcon	There are more apartments in Munich with **maintained condition** compared to Berlin.

Looking at Table 4.1, we can see that Munich has more apartments with a balcony, a kitchen, a higher number of bathrooms, a higher number of bedrooms, a garden, the inclusion of warm water consumption in the heating cost value calculation, a parking space, upscale and normal furnishing, maintained condition, and do not allow pets compared to Berlin but Berlin has more apartments with guest toilets, the inclusion of warm water consumption in the energy consumption value calculation, certificate of eligibility to public housing, and building type of energy performance certificate than Munich.

4.3 Scatter plots of quantitative covariates versus response (rent_sqm) for Berlin and Munich rental properties in 2015 and 2019

Figure 4.3: Scatter plots of quantitative covariates versus response (rent_sqm) with **non linear smooth**: first column = **(rent_sqm)** and second column = **log(rent_sqm)**. (first row) = **Berlin 2015**, (second row)= **Berlin 2019**, (third row) = **Munich 2015**, (fourth row) = **Munich 2019**

The scatter plots of quantitative covariates versus response **rent_sqm**, **log(rent_sqm)**, **log(log(rent_sqm))**, and **1/rent_sqm** with **linear smooth** can be found in Figure A.9 and A.10 of Appendix A

Interpretation of main effects for the quantitative covariates

Table 4.2: Main effects for the quantitative covariates on **rent_sqm**: first block = **rent_sqm**, second block = **log(rent_sqm)**, third block = **log(log(rent_sqm))**, fourth block = **1/rent_sqm**, fifth block = **rent_sqm for non-linear covariates**, sixth block = **log(rent_sqm) for non-linear covariates**

Variables	Berlin 2015	Berlin 2019	Munich 2015	Munich 2019
addcost	Linear(increasing)	Linear (increasing)	Linear (increasing)	nearly constant
heatcost	Linear (increasing)	Linear (increasing)	nearly constant	Linear (decreasing
conyear	constant	constant	constant	constant
lmod	Linear (increasing)	Linear (increasing)	nearly constant	Linear (increasing)
lspace	Linear (increasing)	constant	linear (decreasing)	Linear (decreasing)
fspace	Linear (increasing)	Linear (decreasing)	nearly constant	nearly constant
energycon	Linear (decreasing)	constant	nearly constant	Linear (decreasing)
adlength	Linear (increasing)	Linear (increasing)	Linear (increasing)	Linear (increasing)
addcost	Linear (increasing)	Linear (increasing)	Linear (increasing)	nearly constant
heatcost	Linear (increasing)	Linear (increasing)	constant	Linear (decreasing)
conyear	constant	constant	constant	constant
lmod	Linear (increasing)	Linear (increasing)	Linear (decreasing)	Linear (increasing)
lspace	Linear (increasing)	nearly constant	nearly constant	Linear (decreasing)
fspace	Linear (increasing)	constant	nearly constant	constant
energycon	Linear (decreasing)	constant	nearly constant	nearly constant
adlength	Linear (increasing)	Linear (increasing)	Linear (increasing)	nearly constant
addcost	Linear (increasing)	Linear (increasing)	Linear (increasing)	constant
heatcost	Linear (increasing)	Linear (increasing)	constant	Linear (decreasing)
conyear	nearly constant	constant	constant	constant
lmod	constant	Linear (increasing)	Linear (decreasing)	Linear (increasing)
lspace	Linear (increasing)	nearly constant	constant	Linear (decreasing)
fspace	nearly constant	constant	constant	constant
energycon	nearly constant	constant	constant	constant
adlength	Linear (increasing)	Linear (increasing)	Linear (increasing)	nearly constant
addcost	Linear (decreasing)	Linear (decreasing)	Linear (decreasing)	constant
heatcost	constant	Linear (decreasing)	constant	Linear (increasing)
conyear	Linear (increasing)	nearly constant	constant	constant
lmod	Linear (decreasing)	Linear (decreasing)	Linear (increasing)	constant
lspace	constant	constant	Linear (increasing)	Linear (increasing)
fspace	constant	constant	Linear (increasing)	nearly constant
energycon	nearly constant	constant	nearly constant	constant
adlength	constant	Linear (decreasing)	Linear (decreasing)	constant
addcost	Quadratic	Quadratic	Quadratic	Quadratic
heatcost	nearly linear	Quadratic	Quadratic	Quadratic
conyear	Quadratic	cubic	cubic	Quadratic
lmod	Linear (increasing)	Quadratic	nearly linear	nearly constant
lspace	cubic	cubic	cubic	nearly linear
fspace	cubic	cubic	cubic	Quadratic
energycon	cubic	Quadratic	Quadratic	Quadratic
adlength	Quadratic	cubic	Quadratic	nearly constant
addcost	Quadratic	Quadratic	Quadratic	Quadratic
heatcost	nearly linear	Quadratic	nearly linear	nearly linear
conyear	Quadratic	cubic	Quadratic	nearly linear
lmod	nearly linear	nearly linear	nearly constant	nearly constant
lspace	cubic	cubic	cubic	Linear (decreasing)
fspace	cubic	cubic	cubic	Quadratic
energycon	Quadratic	Quadratic	nearly constant	Quadratic
adlength	cubic	cubic	Quadratic	constant

Looking at the above transformations on **rent_sqm** in Table 4.2, we may likely go with the log transformation for linear and non-linear covariates based on its suitability with respect to constant variance discussed in Chapter 2 and the effects of the covariates on **rent_sqm**.

4.4 Box plots of qualitative covariates versus response (rent_sqm) for Berlin and Munich rental properties in 2015 and 2019

Figure 4.4: Box plots of qualitative covariates versus response (rent_sqm): (first row, first block) = **Berlin 2015**, (second row, first block) = **Berlin 2019**, (first row, second block) = **Munich 2015**, (second row, second block) = **Munich 2019**

Interpretation of main effects for the qualitative covariates

Table 4.3: Main effects for the qualitative covariates on **rent_sqm** in **Berlin 2015**, **Berlin 2019**, **Munich 2015** and **Munich 2019**

Variables	Berlin 2015	Berlin 2019	Munich 2015	Munich 2019
afloorg	Yes	Yes	No	Yes
bfloorg	Yes	Yes	Yes	Yes
nroomsg	Yes	Yes	Yes	Yes
nbedg	Yes	No	No	Yes
nbathg	Yes	Yes	No	No
elevatorg	Yes	Yes	Yes	Yes
balconyg	No	No	No	No
kitcheng	Yes	Yes	Yes	Yes
ewwg	Yes	No	No	No
subhg	Yes	Yes	Yes	Yes
gtoiletg	Yes	Yes	No	No
gardeng	Yes	Yes	No	No
hwwg	Yes	Yes	No	Yes
cellarg	Yes	Yes	Yes	Yes
parkingg	Yes	Yes	No	Yes
furnishingg	Yes	Yes	Yes	Yes
eeffgg	Yes	Yes	Yes	Yes
ecertg	Yes	Yes	No	No
petsg	Yes	Yes	Yes	No
heatg	Yes	Yes	Yes	Yes
apcatg	Yes	Yes	No	No
pcong	Yes	Yes	Yes	No

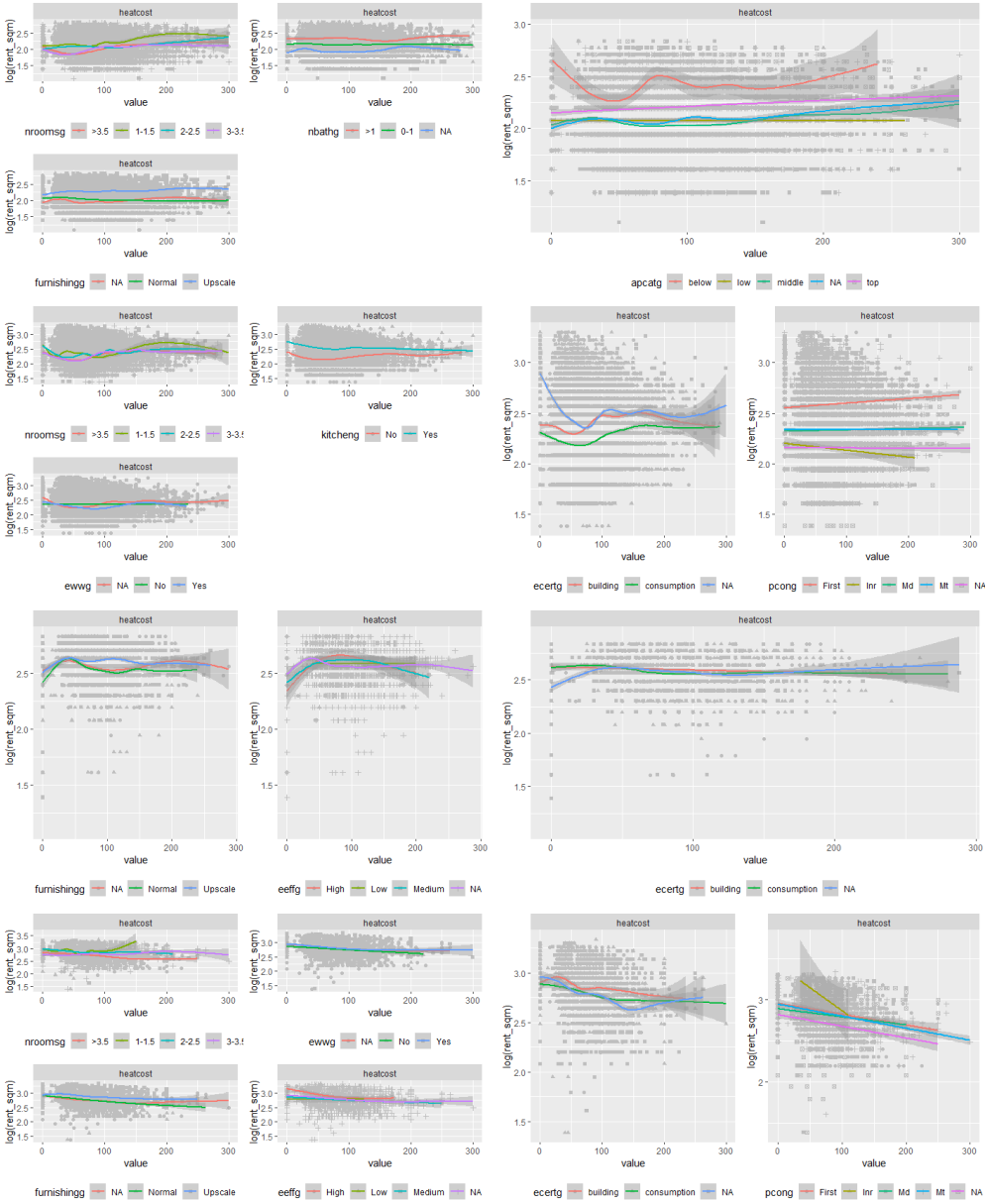
4.5 Interaction effect of heatcost, addcost, covariates and qualitative covariates on rent_sqm for Berlin and Munich rental properties in 2015 and 2019

We would like to study the interaction effects of the below covariates on rent_sqm for Berlin and Munich rental properties in 2015 and 2019

- heatcost and qualitative covariates for non linear smooth on log(rent_sqm)
- addcost and qualitative covariates for non linear smooth on log(rent_sqm)

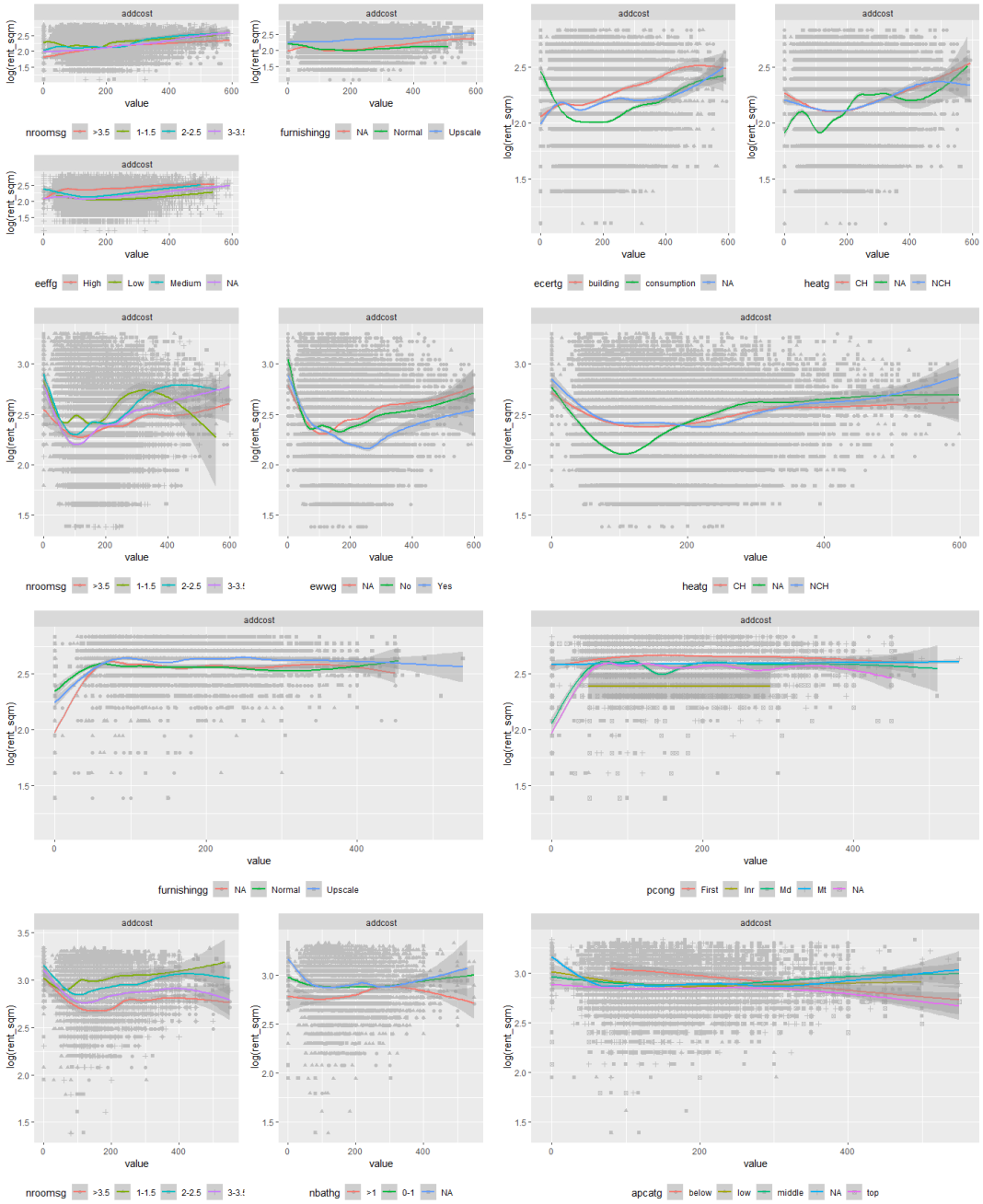
Non linear smooth interaction effect of heatcost with qualitative covariates on log(rent_sqm) for Berlin and Munich in 2015 and 2019

Figure 4.5: Scatter plots of **heatcost** versus **log(rent_sqm)** with **interaction effect** of **qualitative covariates** for **Berlin and Munich in 2015 and 2019**: first row = **Berlin 2015**, second row = **Berlin 2019**, third row = **Munich 2015**, fourth row = **Berlin 2019**



Non linear interaction effect of addcost with qualitative covariates on log(rent_sqm) for Berlin and Munich in 2015 and 2019

Figure 4.6: Scatter plots of **addcost** versus **log(rent_sqm)** with **interaction effect** of **qualitative covariates** for **Berlin and Munich in 2015 and 2019**: first row = **Berlin 2015**, second row = **Berlin 2019**, third row = **Munich 2015**, fourth row = **Berlin 2019**



Interpretation of non linear smooth interaction effects for heatcost, addcost with the qualitative covariates on log(rent_sqm) for Berlin and Munich rental properties in 2015 and 2019

Table 4.4: Non linear smooth interaction effects for heatcost, addcost with the qualitative covariates on **log(rent_sqm)** in **Berlin 2015**, **Berlin 2019**, **Munich 2015** and **Munich 2019**

Quantitative variables	Qualitative variables	Berlin 2015	Berlin 2019	Munich 2015	Munich 2019
	afloorg	No	No	No	No
	bfloorg	No	No	No	No
	nroomsg	Yes	Yes	No	Yes
	nbedg	No	No	No	No
	nbathg	Yes	No	No	No
	elevatorg	No	No	No	No
	balconyg	No	No	No	No
	kitcheng	No	Yes	No	No
	ewwg	No	Yes	No	Yes
	subhg	No	No	No	No
heatcost	gtoiletg	No	No	No	No
	gardeng	No	No	No	No
	hwwg	No	No	No	No
	cellarg	No	No	No	No
	parkingg	No	No	No	No
	furnishingg	Yes	No	Yes	Yes
	eeffgg	No	No	Yes	Yes
	ecertg	No	Yes	Yes	Yes
	petsg	No	No	No	No
	heatg	No	No	No	No
	apcatg	Yes	No	No	No
	pcong	No	Yes	No	Yes
	afloorg	No	No	No	No
	bfloorg	No	No	No	No
	nroomsg	Yes	Yes	No	Yes
	nbedg	No	No	No	No
	nbathg	No	No	No	Yes
	elevatorg	No	No	No	No
	balconyg	No	No	No	No
	kitcheng	No	No	No	No
	ewwg	No	Yes	No	No
	subhg	No	No	No	No
addcost	gtoiletg	No	No	No	No
	gardeng	No	No	No	No
	hwwg	No	No	No	No
	cellarg	No	No	No	No
	parkingg	No	No	No	No
	furnishingg	Yes	No	Yes	No
	eeffgg	Yes	No	No	No
	ecertg	Yes	No	No	No
	petsg	No	No	No	No
	heatg	Yes	Yes	No	No
	apcatg	No	No	No	Yes
	pcong	No	No	Yes	No

Chapter 5

Models fittings and predictions

In this chapter, we discuss how we select the type of model we use to fit the **rent_sqm** for Berlin and Munich rental properties in 2015 and 2019. We first fit the **log(rent_sqm)** on **non linear** covariates without interaction effects (**main effect only**) for **Berlin 2015**, **Berlin 2019**, **Munich 2015** and **Munich 2019**. Thereafter, we fit the **log(rent_sqm)** on **non linear** covariates with the following interaction terms:

- heatcost and qualitative covariates
- addcost and qualitative covariates

5.1 Model type selection

We first fit four models for the response variable in Berlin 2015 in the following cases:

- **Case 1**: We fit a linear regression model where we do not transform the response variable against the covariates (**lm(rent_sqm ~ addcost + heatcost + conyear + lmod + lspace + energycon + adlength + afloorg + bfloorg + nroomsg + nbathg + elevatorg + kitcheng + ewwg + subhg + gtoiletg + gardeng + hwwg + cellarg + parkingg + furnishingg + eeffg + ecertg + petsg + heatg + apcatg + pcong, data = db5_fit**).
- **Case 2**: We fit the log of the response variable against the covariates (**lm(log(rent_sqm) ~ addcost + heatcost + conyear + lmod + lspace + energycon + adlength + afloorg + bfloorg + nroomsg + nbathg + elevatorg + kitcheng + ewwg + subhg + gtoiletg + gardeng + hwwg + cellarg + parkingg + furnishingg + eeffg + ecertg + petsg + heatg + apcatg + pcong, data = db5_fit**).
- **Case 3**: We include a non-linear covariates against the response variable (**lm(rent_sqm ~ poly(addcost,2) + heatcost + poly(conyear,3) + lmod + poly(lspace,3) + poly(fspace,3) + poly(energycon,3) + poly(adlength,2) + afloorg + bfloorg +**

nroomsg + nbathg + elevatorg + kitcheng + ewwg + subhg + gtoiletg + gardeng + hwwg + cellarg + parkingg + furnishingg + eeffg + ecertg + petsg + heatg + apcatg + pcong, data = db5_fit).

- **Case 4**: We include a non-linear covariates against the log of the response variable (**lm(log(rent_sqm) ∼ poly(addcost,2) + heatcost + poly(conyear,3) + lmod + poly(lspace,3) + poly(fspace,3) + poly(energycon,3) + poly(adlength,2) + afloorg + bfloorg + nroomsg + nbathg + elevatorg + kitcheng + ewwg + subhg + gtoiletg + gardeng + hwwg + cellarg + parkingg + furnishingg + eeffg + ecertg + petsg + heatg + apcatg + pcong, data = db5_fit**).

We also do similar model fitting (the 4 cases) for Berlin 2019, Munich 2015, and Munich 2019. The summaries are found in **Table 5.1**.

Table 5.1: Model fitting summary with only main effect

Berlin 2015	**case 1**	**case 2**	**case 3**	**case 4**
Adjusted R-square	0.3081	0.3255	0.3534	0.3645
Number of parameters (p)	42	42	50	52
Berlin 2019				
Adjusted R-square	0.3918	0.4218	0.353	0.4916
Number of parameters (p)	41	41	52	48
Munich 2015				
Adjusted R-square	0.2762	0.2652	0.3101	0.2879
Number of parameters (p)	38	38	39	33
Munich 2019				
Adjusted R-square	0.5139	0.5145	0.3078	0.5468
Number of parameters (p)	25	22	41	27

Looking at the model fitting summary in Table 5.1, we decided to go with the log transformation on **rent_sqm** (**log(rent_sqm)**) for the **non-linear covariates** as it relatively satisfied most of the listed assumptions with larger R-square, compared to the others in the four data sets.

5.2 Model fitting with only main effect on the response

Table 5.2: Model fitting of **log(rent_sqm)** on **non linear** covariates for **Berlin 2015**

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-15.3309	1.6522	-9.28	0.0000
poly(addcost, 2)1	-0.0359	0.2912	-0.12	0.9018
poly(addcost, 2)2	0.8549	0.2182	3.92	0.0001
poly(conyear, 2)1	-2.4303	0.2424	-10.03	0.0000
poly(conyear, 2)2	1.6659	0.2076	8.03	0.0000
lmod	0.0087	0.0008	10.61	0.0000
poly(lspace, 3)1	-1.2691	0.5538	-2.29	0.0220
poly(lspace, 3)2	2.0279	0.3350	6.05	0.0000
poly(lspace, 3)3	-1.2145	0.2288	-5.31	0.0000
poly(fspace, 2)1	-0.7175	0.3432	-2.09	0.0367
poly(fspace, 2)2	-0.3325	0.3048	-1.09	0.2753
bfloorg0-2	0.0061	0.0192	0.32	0.7517
bfloorg3	-0.0014	0.0163	-0.09	0.9318
bfloorg4	0.0490	0.0142	3.45	0.0006
bfloorg5	0.0614	0.0137	4.48	0.0000
bfloorgNA	-0.0013	0.0238	-0.06	0.9559
nroomsg1-1.5	0.0007	0.0284	0.03	0.9798
nroomsg2-2.5	0.0507	0.0224	2.26	0.0237
nroomsg3-3.5	0.0591	0.0176	3.35	0.0008
nbedg0-1	-0.0556	0.0158	-3.52	0.0004
nbedg2	-0.0482	0.0146	-3.31	0.0010
nbedgNA	-0.0689	0.0208	-3.31	0.0009
nbathg0-1	-0.0138	0.0194	-0.71	0.4767
nbathgNA	-0.0626	0.0290	-2.15	0.0313
elevatorgNo	-0.0414	0.0824	-0.50	0.6149
elevatorgYes	0.0096	0.0823	0.12	0.9069
kitchengNo	-0.0378	0.0832	-0.45	0.6493
kitchengYes	0.0315	0.0832	0.38	0.7051
ewwgNo	0.0517	0.0338	1.53	0.1266
ewwgYes	0.0683	0.0342	2.00	0.0455
subhgNo	0.0096	0.0202	0.48	0.6339
subhgYes	-0.2679	0.0482	-5.55	0.0000
gtoiletgYes	0.0352	0.0152	2.32	0.0207
gardengNo	0.1277	0.0659	1.94	0.0527
gardengYes	0.1206	0.0662	1.82	0.0685
parkinggYes	-0.0862	0.0512	-1.68	0.0922
furnishinggNormal	0.0108	0.0147	0.73	0.4643
furnishinggUpscale	0.1274	0.0149	8.54	0.0000
eeffgLow	-0.0753	0.0478	-1.58	0.1152
eeffgMedium	-0.0580	0.0322	-1.80	0.0716
eeffgNA	-0.0189	0.0287	-0.66	0.5102
petsgNo	-0.0377	0.0114	-3.30	0.0010
petsgYes	-0.0479	0.0302	-1.58	0.1132
heatgNA	0.0267	0.0254	1.05	0.2946
heatgNCH	-0.0149	0.0087	-1.72	0.0856
apcatglow	-0.1461	0.0813	-1.80	0.0726
apcatgmiddle	-0.1150	0.0792	-1.45	0.1465
apcatgNA	-0.1284	0.0795	-1.61	0.1064
apcatgtop	-0.0693	0.0795	-0.87	0.3834
pcongInr	-0.1913	0.0615	-3.11	0.0019
pcongMd	-0.0606	0.0100	-6.04	0.0000
pcongMt	-0.0291	0.0117	-2.48	0.0131
pcongNA	-0.0155	0.0205	-0.76	0.4485
Observations	2,605			
R^2	0.377			
Adjusted R^2	0.365			
Residual Std. Error	0.188 (df = 2552)			
F Statistic	29.728*** (df = 52; 2552)			
p-value:	< 2.2e-16			
Note:	*p<0.1; **p<0.05; ***p<0.01			

Table 5.3: Model fitting of **log(rent_sqm)** on **non linear** covariates for **Berlin 2019**

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-13.2034	2.1819	-6.05	0.0000
poly(addcost, 2)1	0.9758	0.3201	3.05	0.0023
poly(addcost, 2)2	-0.2808	0.2557	-1.10	0.2724
heatcost	0.0009	0.0002	4.41	0.0000
poly(conyear, 2)1	-1.2329	0.2958	-4.17	0.0000
poly(conyear, 2)2	2.7830	0.2552	10.90	0.0000
lmod	0.0077	0.0011	7.10	0.0000
poly(lspace, 3)1	-2.7888	0.5671	-4.92	0.0000
poly(lspace, 3)2	1.4259	0.3374	4.23	0.0000
poly(lspace, 3)3	-1.3228	0.2577	-5.13	0.0000
poly(energycon, 2)1	0.3285	0.2503	1.31	0.1896
poly(energycon, 2)2	0.5836	0.2260	2.58	0.0099
poly(adlength, 3)1	0.0461	0.2257	0.20	0.8383
poly(adlength, 3)2	-0.7181	0.2263	-3.17	0.0015
poly(adlength, 3)3	0.4056	0.2166	1.87	0.0613
afloorg>9	-0.0672	0.0668	-1.01	0.3143
afloorg1-2	0.0059	0.0230	0.26	0.7982
afloorg3-9	0.0577	0.0237	2.43	0.0151
afloorgNA	-0.1273	0.0483	-2.64	0.0085
bfloorg0-2	0.0576	0.0322	1.79	0.0742
bfloorg3	0.0056	0.0288	0.20	0.8448
bfloorg4	0.0649	0.0242	2.68	0.0075
bfloorg5	0.1216	0.0215	5.65	0.0000
bfloorgNA	0.1357	0.0413	3.28	0.0011
nroomsg1-1.5	-0.0444	0.0449	-0.99	0.3231
nroomsg2-2.5	0.0550	0.0350	1.57	0.1159
nroomsg3-3.5	0.0192	0.0281	0.68	0.4942
elevatorgYes	0.0900	0.0168	5.36	0.0000
kitchengYes	0.1231	0.0139	8.84	0.0000
ewwgNo	-0.0580	0.0175	-3.32	0.0009
ewwgYes	-0.0091	0.0173	-0.53	0.5972
subhgYes	-0.4174	0.1078	-3.87	0.0001
cellargYes	-0.0260	0.0146	-1.78	0.0749
parkinggNo	0.0970	0.0738	1.31	0.1890
parkinggYes	-0.0655	0.0163	-4.02	0.0001
furnishinggNormal	-0.0390	0.0295	-1.32	0.1872
furnishinggUpscale	0.1026	0.0295	3.47	0.0005
petsgNo	-0.0016	0.0205	-0.08	0.9369
petsgYes	-0.0444	0.0149	-2.97	0.0030
heatgNA	0.0011	0.0530	0.02	0.9833
heatgNCH	-0.0573	0.0144	-3.98	0.0001
apcatglow	-0.0641	0.0924	-0.69	0.4875
apcatgmiddle	-0.0268	0.0839	-0.32	0.7494
apcatgNA	-0.0051	0.0856	-0.06	0.9521
apcatgtop	0.0280	0.0839	0.33	0.7386
pcongInr	-0.1959	0.0850	-2.31	0.0213
pcongMd	-0.0292	0.0210	-1.39	0.1650
pcongMt	-0.0562	0.0223	-2.52	0.0119
pcongNA	0.0227	0.0381	0.59	0.5522
Observations	1,231			
R^2	0.511			
Adjusted R^2	0.492			
Residual Std. Error	0.211 (df = 1182))			
F Statistic	25.778*** (df = 48; 1182)			
p-value:	< 2.2e-16			
Note:	*p<0.1; **p<0.05; ***p<0.01			

Table 5.4: Model fitting of **log(rent_sqm)** on **non linear** covariates for **Munich 2015**

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	2.5838	0.0740	34.90	0.0000
poly(conyear, 2)1	-0.5052	0.1259	-4.01	0.0001
poly(conyear, 2)2	0.5825	0.1307	4.46	0.0000
poly(lspace, 3)1	-1.4841	0.2413	-6.15	0.0000
poly(lspace, 3)2	0.2820	0.1631	1.73	0.0842
poly(lspace, 3)3	-0.3785	0.1423	-2.66	0.0080
adlength	0.0066	0.0030	2.19	0.0290
nroomsg1-1.5	-0.0973	0.0317	-3.06	0.0023
nroomsg2-2.5	-0.0877	0.0227	-3.87	0.0001
nroomsg3-3.5	-0.0542	0.0182	-2.97	0.0031
nbedg0-1	0.0638	0.0211	3.02	0.0026
nbedg2	0.0478	0.0197	2.42	0.0157
nbedgNA	0.0977	0.0279	3.50	0.0005
elevatorgYes	0.0378	0.0096	3.93	0.0001
kitchengNo	-0.0606	0.0383	-1.58	0.1141
kitchengYes	-0.0272	0.0395	-0.69	0.4909
ewwgNo	-0.0808	0.0449	-1.80	0.0721
ewwgYes	-0.0976	0.0450	-2.17	0.0304
subhgNo	0.0422	0.0219	1.93	0.0545
gtoiletgYes	0.0268	0.0138	1.94	0.0528
hwwgYes	0.0205	0.0104	1.98	0.0483
furnishinggNormal	-0.0034	0.0185	-0.18	0.8562
furnishinggUpscale	0.0768	0.0182	4.23	0.0000
eeffgLow	0.1780	0.0658	2.70	0.0070
eeffgMedium	0.1156	0.0478	2.42	0.0158
eeffgNA	0.0725	0.0447	1.62	0.1055
petsgNo	-0.0098	0.0103	-0.95	0.3441
petsgYes	-0.0651	0.0316	-2.06	0.0399
heatgNA	0.0706	0.0213	3.31	0.0010
heatgNCH	-0.0052	0.0121	-0.43	0.6675
pcongInr	0.0700	0.1188	0.59	0.5561
pcongMd	-0.0529	0.0156	-3.40	0.0007
pcongMt	-0.0530	0.0161	-3.30	0.0010
pcongNA	0.0270	0.0262	1.03	0.3031
Observations	711			
R^2	0.321			
Adjusted R^2	0.288			
Residual Std. Error	0.116 (df = 677)			
F Statistic	9.698*** (df = 33; 677)			
p-value:	< 2.2e-16			
Note:	*p<0.1; **p<0.05; ***p<0.01			

Table 5.5: Model fitting of **log(rent_sqm)** on **non linear** covariates for **Munich 2019**

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-8.9984	4.7335	-1.90	0.0586
heatcost	0.0004	0.0003	1.47	0.1425
lmod	0.0060	0.0023	2.54	0.0117
poly(lspace, 2)1	-1.2984	0.2104	-6.17	0.0000
poly(lspace, 2)2	0.4809	0.1481	3.25	0.0013
poly(fspace, 2)1	-0.1854	0.1688	-1.10	0.2733
poly(fspace, 2)2	-0.4037	0.1690	-2.39	0.0178
poly(energycon, 2)1	0.0216	0.1557	0.14	0.8901
poly(energycon, 2)2	0.4404	0.1571	2.80	0.0055
bfloorg0-2	-0.0579	0.0340	-1.70	0.0903
bfloorg3	-0.0029	0.0335	-0.09	0.9322
bfloorg4	-0.0181	0.0314	-0.58	0.5648
bfloorg5	0.0491	0.0320	1.53	0.1266
bfloorgNA	-0.0580	0.1068	-0.54	0.5877
kitchengYes	0.0822	0.0230	3.57	0.0004
hwwgYes	0.0551	0.0216	2.55	0.0116
parkinggNo	-0.1027	0.0561	-1.83	0.0686
parkinggYes	-0.0336	0.0198	-1.69	0.0918
furnishinggNormal	-0.0400	0.0398	-1.01	0.3159
furnishinggUpscale	0.1132	0.0385	2.94	0.0036
ecertgconsumption	-0.0471	0.0214	-2.20	0.0288
apcatglow	-0.1004	0.1636	-0.61	0.5400
apcatgmiddle	-0.1517	0.1564	-0.97	0.3332
apcatgNA	-0.2236	0.1587	-1.41	0.1604
apcatgtop	-0.0951	0.1565	-0.61	0.5441
pcongMd	-0.1170	0.0445	-2.63	0.0091
pcongMt	-0.1001	0.0460	-2.18	0.0305
pcongNA	-0.0194	0.0576	-0.34	0.7362
Observations	244			
R^2	0.597			
Adjusted R^2	0.547			
Residual Std. Error	0.137 (df = 216)			
F Statistic	11.859*** (df = 27; 216)			
p-value:	< 2.2e-16			
Note:	*p<0.1; **p<0.05; ***p<0.01			

5.3 Model fitting with main and interaction effect on the response

We will study the interaction effects of the below covariates on log(rent_sqm) for the non linear covariates.

- heatcost and qualitative covariates
- addcost and qualitative covariates

Table 5.6: Model fitting of **log(rent_sqm)** on **non linear** covariates with interaction for **Berlin 2015**

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-15.3797	1.6433	-9.36	0.0000
poly(conyear, 3)1	-2.4677	0.2417	-10.21	0.0000
poly(conyear, 3)2	1.6163	0.2093	7.72	0.0000
poly(conyear, 3)3	0.4533	0.2047	2.21	0.0269
lmod	0.0084	0.0008	10.43	0.0000
poly(lspace, 3)1	-2.4689	0.4900	-5.04	0.0000
poly(lspace, 3)2	1.7319	0.3088	5.61	0.0000
poly(lspace, 3)3	-1.1566	0.2430	-4.76	0.0000
bfloorg0-2	-0.0009	0.0191	-0.05	0.9604
bfloorg3	-0.0124	0.0165	-0.75	0.4516
bfloorg4	0.0395	0.0144	2.75	0.0060
bfloorg5	0.0530	0.0139	3.82	0.0001
bfloorgNA	0.0079	0.0238	0.33	0.7388
nroomsg1-1.5	0.0520	0.0601	0.86	0.3877
nroomsg2-2.5	0.1341	0.0519	2.59	0.0098
nroomsg3-3.5	0.0587	0.0496	1.18	0.2369
nbedg0-1	-0.0469	0.0160	-2.94	0.0033
nbedg2	-0.0456	0.0147	-3.10	0.0019
nbedgNA	-0.0572	0.0208	-2.75	0.0061
nbathg0-1	0.0141	0.0505	0.28	0.7808
nbathgNA	0.0935	0.0640	1.46	0.1440
elevatorgNo	-0.0164	0.0866	-0.19	0.8496
elevatorgYes	0.0344	0.0866	0.40	0.6912
kitchengNo	0.1609	0.1686	0.95	0.3402
kitchengYes	0.1605	0.1688	0.95	0.3419
ewwgNo	0.0617	0.0339	1.82	0.0687
ewwgYes	0.0738	0.0338	2.18	0.0293
subhgNo	0.0041	0.0201	0.20	0.8388
subhgYes	-0.2627	0.0478	-5.50	0.0000
gtoiletgYes	0.0226	0.0154	1.47	0.1424
gardengNo	0.1310	0.0689	1.90	0.0575
gardengYes	0.1210	0.0692	1.75	0.0806
parkinggYes	-0.0819	0.0505	-1.62	0.1054
furnishinggNormal	-0.0494	0.0372	-1.33	0.1839
furnishinggUpscale	0.0844	0.0366	2.31	0.0212
eeffgLow	-0.2109	0.1124	-1.88	0.0607
eeffgMedium	0.0549	0.0686	0.80	0.4243
eeffgNA	-0.0031	0.0559	-0.06	0.9556
petsgNo	-0.0384	0.0113	-3.40	0.0007
petsgYes	-0.0416	0.0301	-1.38	0.1667
heatgNA	-0.0914	0.0557	-1.64	0.1007
heatgNCH	-0.0585	0.0201	-2.91	0.0036
apcatglow	0.1893	0.1643	1.15	0.2494
apcatgmiddle	0.1333	0.1586	0.84	0.4005
apcatgNA	0.1196	0.1599	0.75	0.4544
apcatgtop	0.1797	0.1601	1.12	0.2620
pcongInr	-0.1909	0.0610	-3.13	0.0018
pcongMd	-0.0589	0.0100	-5.87	0.0000
pcongMt	-0.0307	0.0117	-2.62	0.0087
pcongNA	-0.0268	0.0206	-1.30	0.1937
addcost	0.0011	0.0018	0.61	0.5419
heatcost	0.0040	0.0014	2.82	0.0048
ecertgconsumption	0.0533	0.0197	2.71	0.0068
nroomsg1-1.5:heatcost	-0.0009	0.0005	-1.95	0.0511
nroomsg2-2.5:heatcost	-0.0009	0.0003	-2.67	0.0077
nroomsg3-3.5:heatcost	-0.0006	0.0003	-2.07	0.0381
nbathg0-1:heatcost	-0.0002	0.0004	-0.42	0.6780
nbathgNA:heatcost	-0.0019	0.0006	-3.35	0.0008
furnishinggNormal:heatcost	0.0001	0.0004	0.28	0.7818
furnishinggUpscale:heatcost	-0.0010	0.0004	-2.40	0.0167
apcatglow:heatcost	-0.0034	0.0014	-2.48	0.0133
apcatgmiddle:heatcost	-0.0024	0.0013	-1.84	0.0660
apcatgNA:heatcost	-0.0023	0.0013	-1.77	0.0768
apcatgtop:heatcost	-0.0024	0.0013	-1.86	0.0628
nroomsg1-1.5:addcost	0.0002	0.0004	0.39	0.6976
nroomsg2-2.5:addcost	-0.0002	0.0002	-0.81	0.4196
nroomsg3-3.5:addcost	0.0004	0.0002	1.69	0.0903
kitchengNo:addcost	-0.0019	0.0017	-1.13	0.2587
kitchengYes:addcost	-0.0013	0.0017	-0.75	0.4518
furnishinggNormal:addcost	0.0004	0.0003	1.31	0.1891
furnishinggUpscale:addcost	0.0010	0.0003	3.26	0.0011
eeffgLow:addcost	0.0009	0.0007	1.24	0.2145
eeffgMedium:addcost	-0.0010	0.0005	-1.94	0.0530
eeffgNA:addcost	-0.0002	0.0004	-0.42	0.6779
addcost:ecertgconsumption	-0.0004	0.0001	-3.08	0.0021
heatgNA:addcost	0.0011	0.0004	2.72	0.0066
heatgNCH:addcost	0.0004	0.0002	2.42	0.0154
Observations	2,605			
R^2	0.404			
Adjusted R^2	0.386			
Residual Std. Error	0.185 (df = 2528)			
F Statistic	29.728*** (df = 76; 2528)			
p-value:	< 2.2e-16			
Note:	*p<0.1; **p<0.05; ***p<0.01			

Table 5.7: Model fitting of **log(rent_sqm)** on **non linear** covariates with interaction for **Berlin 2019**

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-13.9794	2.2000	-6.35	0.0000
heatcost	0.0022	0.0006	4.03	0.0001
poly(conyear, 2)1	-1.0750	0.2926	-3.67	0.0003
poly(conyear, 2)2	2.4232	0.2566	9.44	0.0000
lmod	0.0080	0.0011	7.48	0.0000
poly(lspace, 3)1	-3.2428	0.5730	-5.66	0.0000
poly(lspace, 3)2	1.9110	0.3551	5.38	0.0000
poly(lspace, 3)3	-1.1222	0.2682	-4.18	0.0000
poly(energycon, 2)1	0.3853	0.2477	1.56	0.1201
poly(energycon, 2)2	0.6070	0.2210	2.75	0.0061
poly(adlength, 3)1	-0.0100	0.2190	-0.05	0.9636
poly(adlength, 3)2	-0.6442	0.2210	-2.91	0.0036
poly(adlength, 3)3	0.3721	0.2111	1.76	0.0782
afloorg>9	-0.0487	0.0648	-0.75	0.4527
afloorg1-2	0.0056	0.0227	0.24	0.8068
afloorg3-9	0.0486	0.0234	2.08	0.0377
afloorgNA	-0.1092	0.0476	-2.29	0.0220
bfloorg0-2	0.0757	0.0315	2.41	0.0163
bfloorg3	0.0029	0.0281	0.10	0.9167
bfloorg4	0.0622	0.0237	2.62	0.0088
bfloorg5	0.1158	0.0213	5.44	0.0000
bfloorgNA	0.1335	0.0411	3.25	0.0012
nroomsg1-1.5	-0.2009	0.0961	-2.09	0.0369
nroomsg2-2.5	-0.1548	0.0854	-1.81	0.0702
nroomsg3-3.5	-0.0925	0.0827	-1.12	0.2632
elevatorgYes	0.0778	0.0168	4.63	0.0000
kitchengYes	0.2436	0.0278	8.76	0.0000
ewwgNo	-0.1075	0.0463	-2.32	0.0204
ewwgYes	-0.0665	0.0471	-1.41	0.1581
subhgYes	-0.4910	0.1052	-4.67	0.0000
gtoiletgYes	0.0479	0.0241	1.98	0.0474
cellargYes	-0.0440	0.0143	-3.06	0.0022
parkinggNo	0.1222	0.0720	1.70	0.0900
parkinggYes	-0.0576	0.0162	-3.57	0.0004
furnishinggNormal	-0.0335	0.0287	-1.17	0.2441
furnishinggUpscale	0.1048	0.0288	3.64	0.0003
ecertgconsumption	0.0575	0.0339	1.70	0.0901
petsgNo	-0.0240	0.0202	-1.19	0.2355
petsgYes	-0.0538	0.0147	-3.67	0.0003
heatgNA	-0.0603	0.0783	-0.77	0.4419
heatgNCH	-0.1572	0.0305	-5.15	0.0000
apcatglow	-0.0255	0.3976	-0.06	0.9488
apcatgmiddle	0.0137	0.3916	0.04	0.9720
apcatgNA	0.1002	0.3915	0.26	0.7980
apcatgtop	0.1589	0.3923	0.41	0.6855
pcongInr	-0.1603	0.1884	-0.85	0.3950
pcongMd	-0.0301	0.0420	-0.72	0.4731
pcongMt	0.0120	0.0456	0.26	0.7922
pcongNA	0.1251	0.0811	1.54	0.1232
addcost	-0.0000	0.0029	-0.01	0.9925
heatcost:nroomsg1-1.5	-0.0013	0.0007	-1.95	0.0519
heatcost:nroomsg2-2.5	0.0013	0.0005	2.56	0.0106
heatcost:nroomsg3-3.5	0.0011	0.0005	2.34	0.0194
heatcost:kitchengYes	-0.0015	0.0003	-4.84	0.0000
heatcost:ewwgNo	-0.0008	0.0005	-1.57	0.1174
heatcost:ewwgYes	0.0013	0.0005	2.61	0.0091
heatcost:ecertgconsumption	-0.0008	0.0004	-2.21	0.0273
heatcost:pcongInr	-0.0008	0.0022	-0.35	0.7237
heatcost:pcongMd	-0.0001	0.0005	-0.25	0.8020
heatcost:pcongMt	-0.0009	0.0005	-1.87	0.0624
heatcost:pcongNA	-0.0014	0.0009	-1.53	0.1268
nroomsg1-1.5:addcost	0.0014	0.0005	2.78	0.0055
nroomsg2-2.5:addcost	0.0005	0.0003	1.54	0.1240
nroomsg3-3.5:addcost	-0.0001	0.0003	-0.29	0.7704
ewwgNo:addcost	0.0010	0.0003	3.11	0.0019
ewwgYes:addcost	-0.0004	0.0003	-1.26	0.2093
heatgNA:addcost	0.0003	0.0004	0.71	0.4770
heatgNCH:addcost	0.0008	0.0002	4.00	0.0001
apcatglow:addcost	-0.0002	0.0030	-0.08	0.9350
apcatgmiddle:addcost	-0.0002	0.0030	-0.07	0.9449
apcatgNA:addcost	-0.0008	0.0030	-0.27	0.7882
apcatgtop:addcost	-0.0009	0.0030	-0.30	0.7654
Observations	1,231			
R^2	0.556			
Adjusted R^2	0.529			
Residual Std. Error	0.203 (df = 1159)			
F Statistic	20.437*** (df = 71; 1159)			
p-value:	< 2.2e-16			
Note:	*p<0.1; **p<0.05; ***p<0.01			

Table 5.8: Model fitting of **log(rent_sqm)** on **non linear** covariates with interaction for **Munich 2015**

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	2.3013	0.1521	15.13	0.0000
heatcost	0.0058	0.0011	5.16	0.0000
poly(conyear, 2)1	-0.5036	0.1322	-3.81	0.0002
poly(conyear, 2)2	0.6351	0.1304	4.87	0.0000
poly(lspace, 3)1	-1.4493	0.2703	-5.36	0.0000
poly(lspace, 3)2	0.2814	0.1646	1.71	0.0877
poly(lspace, 3)3	-0.4669	0.1403	-3.33	0.0009
adlength	0.0067	0.0029	2.29	0.0225
bfloorg0-2	-0.0326	0.0193	-1.69	0.0917
bfloorg3	-0.0301	0.0178	-1.69	0.0917
bfloorg4	-0.0316	0.0175	-1.80	0.0716
bfloorg5	0.0101	0.0172	0.59	0.5566
bfloorgNA	0.0432	0.0287	1.50	0.1329
nroomsg1-1.5	-0.0916	0.0331	-2.77	0.0057
nroomsg2-2.5	-0.0811	0.0244	-3.33	0.0009
nroomsg3-3.5	-0.0690	0.0191	-3.62	0.0003
nbedg0-1	0.0585	0.0224	2.61	0.0093
nbedg2	0.0555	0.0201	2.77	0.0058
nbedgNA	0.0967	0.0336	2.88	0.0041
nbathg0-1	0.0135	0.0429	0.31	0.7530
nbathgNA	-0.1495	0.0795	-1.88	0.0604
elevatorgYes	0.0205	0.0118	1.73	0.0841
kitchengNo	-0.0608	0.0371	-1.64	0.1019
kitchengYes	-0.0217	0.0383	-0.57	0.5709
ewwgNo	-0.1628	0.0945	-1.72	0.0853
ewwgYes	-0.1195	0.0939	-1.27	0.2039
subhgNo	0.0391	0.0211	1.85	0.0651
gtoiletgYes	0.0344	0.0140	2.45	0.0147
hwwgYes	0.0216	0.0121	1.78	0.0759
furnishinggNormal	0.0427	0.0450	0.95	0.3426
furnishinggUpscale	0.0938	0.0443	2.12	0.0344
eeffgLow	-0.5593	0.2520	-2.22	0.0268
eeffgMedium	0.5698	0.1069	5.33	0.0000
eeffgNA	0.5416	0.0959	5.65	0.0000
ecertgconsumption	-0.0894	0.0307	-2.91	0.0037
heatgNA	0.0652	0.0213	3.06	0.0023
heatgNCH	-0.0032	0.0118	-0.27	0.7843
pcongInr	0.0532	0.1132	0.47	0.6387
pcongMd	-0.1456	0.0344	-4.24	0.0000
pcongMt	-0.1310	0.0341	-3.84	0.0001
pcongNA	0.0135	0.0580	0.23	0.8157
addcost	-0.0012	0.0008	-1.61	0.1077
heatcost:furnishinggNormal	-0.0010	0.0005	-2.01	0.0453
heatcost:furnishinggUpscale	-0.0009	0.0005	-1.93	0.0538
heatcost:eeffgLow	0.0077	0.0027	2.86	0.0043
heatcost:eeffgMedium	-0.0055	0.0011	-4.78	0.0000
heatcost:eeffgNA	-0.0057	0.0011	-5.39	0.0000
heatcost:ecertgconsumption	0.0008	0.0003	2.58	0.0100
nbathg0-1:addcost	-0.0002	0.0002	-0.98	0.3286
nbathgNA:addcost	0.0007	0.0004	1.60	0.1108
ewwgNo:addcost	0.0005	0.0006	0.79	0.4285
ewwgYes:addcost	0.0001	0.0006	0.22	0.8275
furnishinggNormal:addcost	0.0005	0.0003	1.50	0.1333
furnishinggUpscale:addcost	0.0007	0.0003	2.00	0.0457
pcongMd:addcost	0.0007	0.0002	2.89	0.0040
pcongMt:addcost	0.0006	0.0002	2.40	0.0167
pcongNA:addcost	0.0001	0.0004	0.25	0.7989
Observations	711			
R^2	0.409			
Adjusted R^2	0.358			
Residual Std. Error	0.110 (df = 654)			
F Statistic	8.068^{***} (df = 56; 654)			
p-value:	< 2.2e-16			
Note:	*p<0.1; **p<0.05; ***p<0.01			

Table 5.9: Model fitting of **log(rent_sqm)** on **non linear** covariates with interaction for **Munich 2019**

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-11.0895	5.6609	-1.96	0.0517
heatcost	0.0057	0.0021	2.72	0.0072
poly(conyear, 2)1	0.1326	0.1779	0.75	0.4570
poly(conyear, 2)2	0.3935	0.1829	2.15	0.0329
lmod	0.0071	0.0028	2.53	0.0122
poly(lspace, 2)1	-1.7236	0.3802	-4.53	0.0000
poly(lspace, 2)2	0.6788	0.2401	2.83	0.0053
poly(energycon, 2)1	0.1392	0.1846	0.75	0.4521
poly(energycon, 2)2	0.3338	0.1783	1.87	0.0629
bfloorg0-2	-0.0784	0.0379	-2.07	0.0403
bfloorg3	0.0367	0.0384	0.95	0.3411
bfloorg4	-0.0229	0.0336	-0.68	0.4953
bfloorg5	0.0532	0.0356	1.50	0.1366
bfloorgNA	-0.0635	0.1063	-0.60	0.5508
nroomsg1-1.5	-0.1532	0.1975	-0.78	0.4389
nroomsg2-2.5	0.1057	0.1758	0.60	0.5486
nroomsg3-3.5	0.2912	0.1764	1.65	0.1007
nbedg0-1	0.0405	0.0507	0.80	0.4260
nbedg2	0.0513	0.0453	1.13	0.2587
nbedgNA	0.1380	0.0729	1.89	0.0600
nbathg0-1	-0.3568	0.1385	-2.58	0.0108
nbathgNA	-0.6844	0.2943	-2.33	0.0212
kitchengYes	0.1346	0.0631	2.13	0.0343
ewwgNo	-0.2206	0.0981	-2.25	0.0258
ewwgYes	0.1099	0.0781	1.41	0.1609
gtoiletgYes	0.0524	0.0291	1.80	0.0739
hwwgYes	0.0555	0.0302	1.84	0.0680
parkinggNo	-0.0226	0.0582	-0.39	0.6988
parkinggYes	-0.0585	0.0216	-2.71	0.0074
furnishinggNormal	0.3206	0.1253	2.56	0.0114
furnishinggUpscale	0.3891	0.1245	3.13	0.0021
eeffgLow	-0.0489	0.2141	-0.23	0.8196
eeffgMedium	0.1443	0.1936	0.75	0.4571
eeffgNA	-0.0796	0.1637	-0.49	0.6274
ecertgconsumption	0.0839	0.0646	1.30	0.1961
petsgNo	0.0394	0.0260	1.52	0.1314
petsgYes	0.1033	0.0299	3.45	0.0007
heatgNA	0.0952	0.0566	1.68	0.0946
heatgNCH	0.0169	0.0290	0.58	0.5603
apcatglow	-0.9409	0.3453	-2.73	0.0071
apcatgmiddle	-0.8986	0.3077	-2.92	0.0040
apcatgNA	-0.8741	0.3123	-2.80	0.0057
apcatgtop	-0.6706	0.2847	-2.36	0.0196
pcongMd	0.1800	0.2023	0.89	0.3750
pcongMt	0.0859	0.1997	0.43	0.6678
pcongNA	-0.1664	0.2505	-0.66	0.5075
addcost	-0.0006	0.0016	-0.36	0.7203
heatcost:nroomsg1-1.5	0.0037	0.0017	2.25	0.0260
heatcost:nroomsg2-2.5	0.0011	0.0011	0.95	0.3451
heatcost:nroomsg3-3.5	0.0001	0.0011	0.13	0.8978
heatcost:ewwgNo	0.0022	0.0010	2.26	0.0251
heatcost:ewwgYes	-0.0005	0.0008	-0.65	0.5156
heatcost:furnishinggNormal	-0.0041	0.0013	-3.11	0.0022
heatcost:furnishinggUpscale	-0.0034	0.0013	-2.64	0.0091
heatcost:eeffgMedium	-0.0020	0.0010	-2.00	0.0468
heatcost:ecertgconsumption	-0.0015	0.0007	-2.13	0.0344
heatcost:pcongMd	-0.0028	0.0010	-2.68	0.0081
heatcost:pcongMt	-0.0013	0.0010	-1.28	0.2031
heatcost:pcongNA	-0.0019	0.0021	-0.91	0.3662
nroomsg1-1.5:addcost	-0.0007	0.0010	-0.69	0.4880
nroomsg2-2.5:addcost	-0.0014	0.0007	-1.89	0.0602
nroomsg3-3.5:addcost	-0.0019	0.0007	-2.73	0.0070
nbathg0-1:addcost	0.0018	0.0008	2.39	0.0181
nbathgNA:addcost	0.0029	0.0018	1.65	0.1005
kitchengYes:addcost	-0.0006	0.0004	-1.32	0.1871
ewwgNo:addcost	0.0006	0.0005	1.20	0.2317
ewwgYes:addcost	-0.0005	0.0005	-1.03	0.3060
apcatglow:addcost	0.0025	0.0009	2.77	0.0063
apcatgmiddle:addcost	0.0013	0.0004	3.05	0.0027
apcatgNA:addcost	0.0006	0.0006	1.04	0.3021
pcongMd:addcost	-0.0006	0.0012	-0.52	0.6029
pcongMt:addcost	-0.0006	0.0012	-0.48	0.6291
pcongNA:addcost	0.0018	0.0014	1.27	0.2061
Observations	244			
R^2	0.720			
Adjusted R^2	0.603			
Residual Std. Error	0.128 (df = 171)			
F Statistic	6.120^{***} (df = 72; 171)			
p-value:	< 2.2e-16			
Note:	*p<0.1; **p<0.05; ***p<0.01			

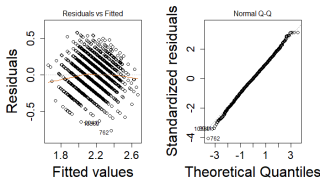
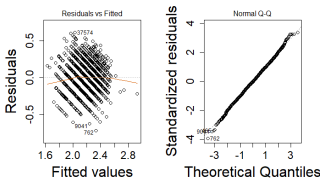
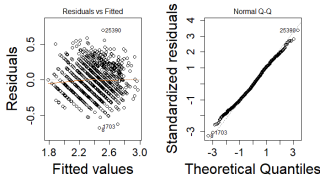
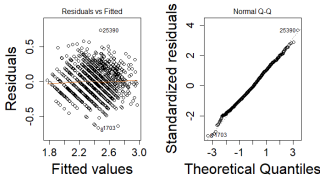
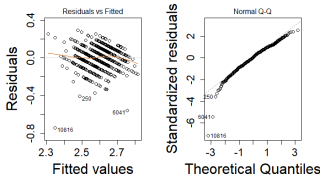
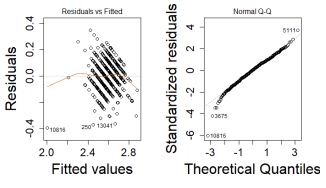
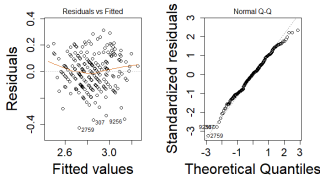
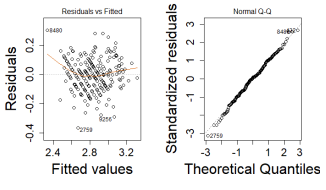
Table 5.10: Summary of the inclusion of the quantitative covariates, the interaction effect of addcost with the qualitative covariates, and the interaction effect of heatcost with the qualitative covariates for both the **main effects and the interaction effect models** in **Berlin 2015**, **Berlin 2019**, **Munich 2015** and **Munich 2019**

Models	Variables	Berlin 2015	Berlin 2019	Munich 2015	Munich 2019
main effect					
	addcost	✓	✓	-	-
	heatcost	-	✓	-	✓
	conyear	✓	✓	✓	-
	lmod	✓	✓	-	✓
	lspace	✓	✓	✓	✓
	fspace	✓	-	-	✓
	energycon	-	✓	-	✓
	adlength	-	✓	✓	-
interaction effect					
	addcost*afloorg	-	-	-	-
	addcost*bfloorg	-	-	-	-
	addcost*nroomsg	✓	✓	-	✓
	addcost*nbedg	-	-	-	-
	addcost*nbathg	-	-	✓	✓
	addcost*elevatorg	-	-	-	-
	addcost*balconyg	-	-	-	-
	addcost*kitcheng	✓	-	-	✓
	addcost*ewwg	-	✓	✓	✓
	addcost*subhg	-	-	-	-
	addcost*gtoiletg	-	-	-	-
	addcost*gardeng	-	-	-	-
	addcost*hwwg	-	-	-	-
	addcost*cellarg	-	-	-	-
	addcost*parkingg	-	-	-	-
	addcost*furnishingg	✓		✓	-
	addcost*eeffgg	✓	-	-	-
	addcost*ecertg	✓	-	-	-
	addcost*petsg	-	-	-	-
	addcost*heatg	✓	✓	-	-
	addcost*apcatg	-	✓	-	✓
	addcost*pcong	-	-	✓	✓
	heatcost*afloorg	-	-	-	-
	heatcost*bfloorg	-	-	-	-
	heatcost*nroomsg	✓	✓	-	✓
	heatcost*nbedg	-	-	-	-
	heatcost*nbathg	✓	-	-	-
	heatcost*elevatorg	-	-	-	-
	heatcost*balconyg	-	-	-	-
	heatcost*kitcheng	-	✓	-	-
	heatcost*ewwg	-	✓	-	✓
	heatcost*subhg	-	-	-	-
	heatcost*gtoiletg	-	-	-	-
	heatcost*gardeng	-	-	-	-
	heatcost*hwwg	-	-	-	-
	heatcost*cellarg	-	-	-	-
	heatcost*parkingg	-	-	-	-
	heatcost*furnishingg	✓	-	✓	✓
	heatcost*eeffgg	-	-	✓	✓
	heatcost*ecertg	-	✓	✓	✓
	heatcost*petsg	-	✓	-	-
	heatcost*heatg	-	-	-	-
	heatcost*apcatg	✓	-	-	-
	heatcost*pcong	-	-	-	✓

5.4 Residual plots of model fittings

In this section, we plot the residuals versus the fitted values to see if there is a trend in order to check for the plausibility of the linearity assumption discussed in Chapter 2. Also, we plot the QQ plots of the covariates versus the theoretical normal quantile to see if it is a straight line in order to check for the plausibility of the normality assumption which was discussed in Chapter 2.

Table 5.11: Residual plots of model fittings for **Berlin 2015**, **Berlin 2019**, **Munich 2015** and **Munich 2019**

city	main effect	interaction effect
Berlin 2015		
Berlin 2019		
Munich 2015		
Munich 2019		

From the plots in Table 5.11, we find that the fitted models relatively do not violate the linear regression assumptions in Chapter 2.

5.5 Comparing models using ANOVA

In this section, we compare the **main effect models without the interaction effects** and the **model with the interaction terms included** to see if the model with interaction terms is significantly better than the model with only the main effects. For Berlin 2015, we compare the two models specified in Table 5.2 and Table 5.6. Similarly, we compare the models specified in Table 5.3 and Table 5.7 for Berlin 2019, and, for Munich 2015, we compare the models specified in Table 5.4 and Table 5.8. Finally, we compare the two models specified in Table 5.5 and Table 5.9 for Munich 2019.

Table 5.12: **ANOVA** analysis of **only main effects** and **main and interaction effects** model fittings for **Berlin 2015**, **Berlin 2019**, **Munich 2015** and **Munich 2019**

city	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
Berlin 2015	2552	90.51				
	2528	86.62	24	3.89	4.73	0.0000
Berlin 2019	1182	52.59				
	1159	47.80	23	4.79	5.05	0.0000
Munich 2015	677	9.14				
	654	7.96	23	1.18	4.21	0.0000
Munich 2019	216	4.05				
	171	2.81	45	1.24	1.68	0.0101

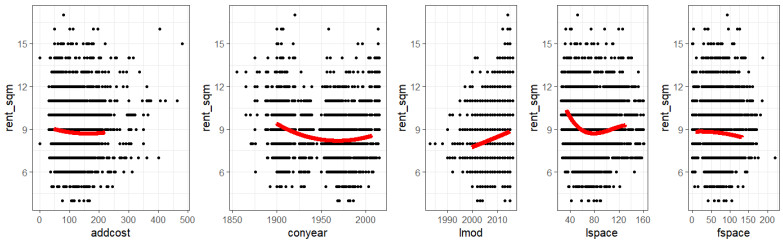
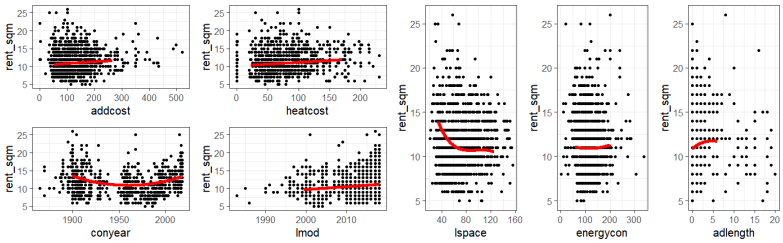
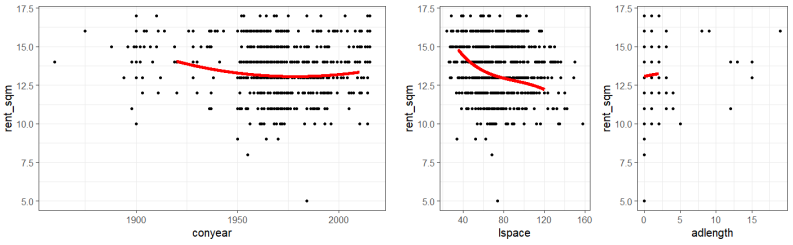
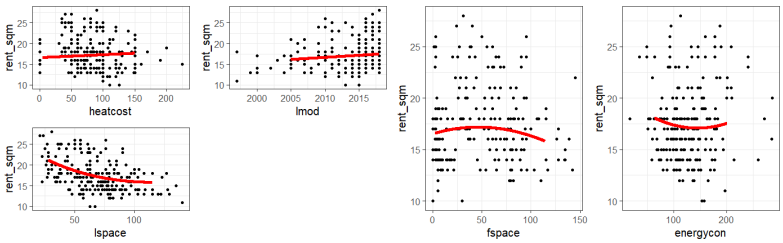
From the result in Table 5.12, we can see that the interaction effects are significant and therefore better than the models with only main effects.

5.6 Model predictions of rent_sqm for the main effect models

In this section, we will predict the values of **rent_sqm** for the main effect models given in Table 5.2, Table 5.3, Table 5.4, and Table 5.5. We will use the **median** of the continuous covariates and the **mode** of the qualitative covariates for our prediction. We consider the mode for the qualitative covariates and the median for the remaining continuous variables while we take 50 values between the 5th and 95th quantile/percentile of the variable we

are plotting. We also consider the different categories of each qualitative covariate which we are using for the prediction of **rent_sqm** while other qualitative covariates remain in their mode and the continuous covariates in their medians respectively.

Table 5.13: Model predictions of **rent_sqm** for the **quantitative covariates** in **Berlin 2015**, **Berlin 2019**, **Munich 2015** and **Munich 2019 main effect models**

city	prediction plots
Berlin 2015	
Berlin 2019	
Munich 2015	
Munich 2019	

Interpretation of model predictions of rent_sqm for the quantitative covariates in Table 5.13

From Table 5.13, we can summarise the behaviour of the **predicted rent_sqm** as follows

- We can see that each of the variables in **Berlin 2015** has an influence on **rent_sqm** as their predicted lines are not constant. The last modernization variable enters the model linearly and has an increasing trend with the **rent_sqm**, unlike the other variables which do not enter the model linearly.
- In **Berlin 2019**, each of the variables also has an influence on the **rent_sqm**. The additional cost, heat cost, and last modernization variables enter the model linearly and have an increasing trend with the **rent_sqm** while the other variables enter the model nonlinearly.
- In **Munich 2015**, we also see that all the variables have an influence on **rent_sqm**. The length of advertisement enters the model linearly and has an increasing trend with the **rent_sqm** while the construction year and living space variables enter the model nonlinearly, although we see a decreasing trend in the living space with the **rent_sqm**.
- Finally, in **Munich 2019**, all the variables equally have an influence on **rent_sqm**. The heat cost and the last modernization variables enter the model linearly and have an increasing trend with the **rent_sqm** while the other variables enter the model nonlinearly, although we see a decreasing trend in the living space with the **rent_sqm**.

Table 5.14: Model predictions of **rent_sqm** for the **qualitative covariates** in **Berlin 2015**, **Berlin 2019**, **Munich 2015** and **Munich 2019 main effect models**

Variables	categories	Berlin 2015	Berlin 2019	Munich 2015	Munich 2019
afloorg	(-1)-0		10.30		
	1-2		10.36		
	3-9		10.91 (mode = 3-9)		
	> 9		9.63		
	NA		9.07		
bfloorg	0-2	8.41	10.83		16.50
	3	8.34	10.29		17.44
	4	8.78 (mode = 4)	10.91 (mode = 4)		17.17 (mode = 4)
	5	8.89	11.55		18.37
	>5	8.36	10.23		17.49
	NA	8.35	11.72		16.50
nroomsg	1-1.5	8.35	9.88	12.96	
	2-2.5	8.78 (mode = 2-2.5)	10.91 (mode = 2-2.5)	13.09 (mode = 2-2.5)	
	3-3.5	8.85	10.53	13.53	
	> 3.5	8.34	10.33	14.29	
nbedg	0-1	8.78 (mode = 0-1)		13.09 (mode = 0-1)	
	2	8.84		12.88	
	> 2	9.28		12.28	
	NA	8.66		13.54	
nbathg	0-1	8.78 (mode = 0-1)			
	> 1	8.90			
	NA	8.36			
elevatorg	Yes	9.24	11.94	13.59	
	No	8.78 (mode = No)	10.91 (mode = No)	13.09 (mode = No)	
	NA	9.15			
balconyg					
kitcheng	Yes	9.41	10.91 (mode = Yes)	13.09 (mode = Yes)	17.17 (mode = Yes)
	No	8.78 (mode = No)	9.65	12.66	15.82
	NA	9.11		13.45	
ewwg	Yes	8.92	10.81	12.87	
	No	8.78 (mode = No)	10.30	13.09 (mode = No)	
	NA	8.33	10.91 (mode = NA)	14.19	
subhg	Yes	6.65	7.19		
	No	8.78 (mode = No)	10.91 (mode = No)	13.09 (mode = No)	
	NA	8.69		12.55	
gtoiletg	Yes	9.09		13.44	
	No	8.78 (mode = No)		13.09 (mode = No)	
gardeng	Yes	8.71			
	No	8.78 (mode = No)			
	NA	7.72			
hwwg	Yes			13.36	18.14
	No			13.09 (mode = No)	17.17 (mode = No)
cellarg	Yes		10.91 (mode = Yes)		
	No		11.20		
parkingg	Yes	8.05	10.22		17.17 (mode = Yes)
	No		12.03		16.02
	NA	8.78 (mode = NA)	10.91 (mode = NA)		17.77
furnishingg	Upscale	8.78 (mode = Upscale)	10.91 (mode = Upscale)	13.09 (mode = Upscale)	17.17 (mode = Upscale)
	Normal	7.81	9.47	12.08	14.73
	NA	7.73	9.85	12.12	15.33
eeffgg	High	8.94		12.17	
	Meduim	8.44		13.66	
	Low	8.30		14.54	
	NA	8.78 (mode = NA)		13.087 (mode = NA)	
ecertg	consumption				17.17 (mode = consumption)
	building				18.00
petsg	Yes	8.37	10.91 (mode = Yes)	12.26	
	No	8.45	11.39	12.96	
	NA	8.78 (mode = NA)	11.41	13.09 (mode = NA)	
heatg	CH	8.78 (mode = CH)	10.91 (mode = CH)	13.09 (mode = CH)	
	NCH	8.65	10.31	13.02	
	NA	9.01	10.93	14.04	
apcatg	top	9.19	11.53		18.17
	middle	8.78 (mode = middle)	10.91 (mode = middle)		17.17 (mode = middle)
	low	8.51	10.51		18.08
	below	9.85	11.21		19.99
	NA	8.67	11.15		15.98
pcong	Md	8.78 (mode = Md)	10.91 (mode = Md)	13.09 (mode = Md)	17.17 (mode = Md)
	Mt	9.06	10.62	13.09	17.46
	First	9.32	11.24	13.80	19.30
	Inr	7.70	9.24	14.80	
	NA	9.18	11.50	14.18	18.93

Interpretation of model predictions of rent_sqm for the qualitative covariates in Table 5.14

We can summarise the behaviour of the **predicted rent_sqm** in Table 5.14 as follows

- The predicted **rent_sqm** is relatively **higher in Munich** compared to **Berlin** considering the predicted values with their respective mode categories.
- **afloor:** In Berlin 2019, the predicted **rent_sqm** is at the lowest (9.63 Euros) with apartment floor >9. It increased with apartment floor (-1)-0 (10.30 Euros), followed by apartment floor 1-2 (10.36 Euros) and it is at the highest (10.91 Euros) with apartment floor 3-9 (the mode).
- **building floors (bfloor):** With 5 building floors apartments, our predicted **rent_sqm** is at the highest (8.89, 11.55, and 18.37 Euros) for Berlin 2015, Berlin 2019, and Munich 2019 respectively. Also, the predicted **rent_sqm** is at the lowest (8.34 and 10.29 Euros) with 3 building floors apartments in Berlin 2015 and 2019 respectively but in Munich 2019, it is at the lowest with building floors apartments 0-2 (16.50 Euros).
- **number of rooms (nrooms):** In **Berlin 2015**, our **predicted rent_sqm** value is at the highest (8.85 Euros) with apartments that have **the number of rooms 3-3.5** while in **Berlin 2019**, the apartments with **the number of rooms 2-2.5** has the highest **predicted rent_sqm** (10.91 Euros). Also, the predicted **rent_sqm** is at the lowest with apartments that have 1-1.5 (8.35, 9.88, and 12.96 Euros) for Berlin 2015, Berlin 2019, and Munich 2015 respectively but it is at the highest in Munich 2015 with apartments that have ¿3.5 rooms (14.29 Euros).
- **number of bedrooms (nbed)**: In **Berlin 2015**, we can see **an increasing trend** in the **predicted rent_sqm** (8.78, 8.84, and 9.28 Euros) with respect to the order of the categories of the **number of bedrooms (0-1, 2, and > 2)** of an apartment while in **Munich 2015**, the reverse is the case as we can see **a decreasing trend** in the **predicted rent_sqm** (13.09, 12.88, and 12.28 Euros) with respect to the same order of the categories of the **number of bedrooms (0-1, 2, and > 2)**. Thus rent_sqm increases with apartments that have higher number of bedrooms in Berlin but decreases in Munich (vice versa).
- **number of bathrooms (nbath):** In Berlin 2015, we can also see **an increasing trend** in the **predicted rent_sqm** (8.78 and 8.90 Euros) with respect to the order of the categories of the **number of bathrooms (0-1, and > 1)** of an apartment. Thus rent_sqm increases with apartments that have higher number of bathrooms in Munich (vice versa).
- **elevator:** We can see **an increase** in the **predicted rent_sqm** (9.24, 11.94, and 13.59 Euros) for apartments with an elevator in Berlin 2015, Berlin 2019 and Munich 2015 respectively unlike the apartments without an elevator where the **predicted rent_sqm** are respectively (8.78, 10.91, and 13.09 Euros). Thus rent_sqm increases with apartments that have an elevator (vice versa).
- **kitchen:** We can see **an increase** in the **predicted rent_sqm** (9.41, 10.91, 13.09, and 17.17 Euros) for apartments with a kitchen in Berlin 2015, Berlin 2019, Munich 2015 and

Munich 2019 respectively unlike the apartments without a kitchen where the **predicted rent_sqm** are respectively (8.78, 9.65, 12.66, and 15.82 Euros). Thus rent_sqm increases with apartments that have a kitchen (vice versa).

- **eww:** The **predicted rent_sqm** is higher with apartments that have the inclusion of warm water consumption in the energy consumption value calculation (8.92 and 10.81 Euros) in both Berlin 2015 and 2019 respectively, compared to the apartments that do not have it (8.78 and 10.30 euros). On the other hand, the **predicted rent_sqm** in Munich 2019 is lower with apartments that have it (12.87 Euros) compared with the apartments that do not have it (13.09 Euros). Thus rent_sqm increases with apartments that have the inclusion of warm water consumption in the energy consumption value calculation in Berlin (vice versa) but decreases in Munich with apartments that have it (vice versa).

- **subh:** The **predicted rent_sqm** is lower with apartments that have a certificate of eligibility to public housing (6.65 and 7.19 Euros) in both Berlin 2015 and 2019 respectively, compared to the apartments that do not have it (8.78 and 10.91 Euros). Thus rent_sqm increases with apartments that do not have a certificate of eligibility to public housing in Berlin (vice versa).

- **gtoilet:** The **predicted rent_sqm** is higher with apartments that have guest toilet (9.09 and 13.44 Euros) in both Berlin 2015 and Munich 2015 respectively, compared to the apartments with no guest toilet (8.78 and 13.08 Euros). Thus rent_sqm increases with apartments that have guest toilet in both Berlin and Munich (vice versa).

- **garden:** With apartments that have garden in Berlin 2015, the **predicted rent_sqm** is lower (8.71 Euros) compared to apartments that do not have garden (8.78 Euros).

- **hww:** With apartments that have the warm water consumption included in the heating cost value calculation in both Munich 2015 and 2019, the **predicted rent_sqm** is higher (13.36 and 18.14 Euros) compared to apartments that do not have it (13.09 and 17.17 Euros), thereby increased by 36% in Munich from 2015 to 2019 with apartments that have the warm water consumption included in the heating cost value calculation. Thus rent_sqm increases with apartments that have the warm water consumption included in the heating cost value calculation in Munich (vice versa)

- **cellar:** With apartments that have a cellar in Berlin 2019, the **predicted rent_sqm** is lower (10.91 Euros) compared to apartments that do not have cellars (11.20 Euros). Thus rent_sqm decreases with apartments that have a cellar in Berlin (vice versa)

- **parking space:** In Berlin 2019, the **predicted rent_sqm** is higher (10.22 Euros) with apartments that have a parking space, compared to apartments that do not have a parking space (12.03 Euros). In Munich 2019, the **predicted rent_sqm** is also higher (17.17 Euros) with apartments that have a parking space, compared to apartments that do not have a parking space (16.02 Euros). Thus rent_sqm increases in both Berlin and Munich with apartments that have a parking space (vice versa)

- **furnishing:** The predicted **rent_sqm** is at the highest with apartments that have **Upscale furnishing** for **Berlin 2015**, **Berlin 2019**, **Munich 2015**, and **Munich 2019**. Also, with

an Upscale furnishing apartments, the predicted **rent_sqm** increased from 2015 to 2019 by 24.26% and 31.17% for Berlin and Munich respectively. It equally increased from Normal to Upscale furnishing apartments by 12.42%, 15.21%, 8.36%, and 16.56% for Berlin 2015, Berlin 2019, Munich 205 and Munich 2019 respectively. Thus, rent_sqm increases with apartments that have Upscale furnishing in both Berlin and Munich (vice versa), as well as with respect to time.

- **energy efficiency rating (eeff)**: We can also see **an increasing trend** in the **predicted rent_sqm** with respect to the order of the categories of **energy efficiency rating (Low, Medium, and High)** (8.30, 8.44, and 8.94 Euros) in **Berlin 2015** while in **Munich 2015**, the reverse is the case as we can see **a decreasing trend** (14.54, 13.67, and 12.17 Euros). Thus rent_sqm increases in Berlin but decreases in Munich with respect to the order of energy efficiency rating categories (Low, Medium, and High) (vice versa).

- **ecertg:** In Munich 2019, the **predicted rent_sqm** is higher with apartments that have the **building type of energy performance certificate** (18.00 Euros) compared to apartments that have the **construction type of energy performance certificate** (17.17 Euros).

- **pets:** The **predicted rent_sqm** is lower with apartments that allow pets (8.37, 10.91, and 12.26 Euros) in Berlin 2015, Berlin 2019, and Munich 2015 respectively, compared to the apartments that do not allow pets (8.45, 11.39, and 12.96 Euros), thereby decreased by 1%, 4% and 5% for Berlin 2015, Berlin 2019, and Munich 2015 respectively. Thus, rent_sqm increases with apartments that do not allow pets in both Berlin and Munich (vice versa)

- **heat:** Our predicted **rent_sqm** is higher with apartments that make use of the **central heating (CH)** as their heating type (8.78, 10.91, and 13.09 Euros) in Berlin 2015, Berlin 2019, and Munich 2015 respectively, compared to the apartments that make use of the **non-central heating (NCH)** as their heating type (8.65, 10.31, and 13.02 Euros), thereby increased by 2%, 6% and 1% for Berlin 2015, Berlin 2019, and Munich 2015 respectively. Thus, rent_sqm increases with apartments that make use of the central heating as their heating type in both Berlin and Munich (vice versa)

- **apartment categories (apcat)**: Our predicted **rent_sqm** is at the highest with the **below category** apartments (9.85 and 19.99 Euros) for **Berlin 2015** and Munich 2019 respectively, while for **Berlin 2019**, it is at the highest with the **top category** apartments (11.53 Euros). On the other hand, our predicted **rent_sqm** is at the lowest with the **low category** apartments (8.51 and 10.51 Euros) for **Berlin 2015** and 2019 respectively but in Munich 2019 it is at the lowest with the **middle category** apartments (17.17 Euros).

- **property condition categories (pcon)**: Our predicted **rent_sqm** is at the highest with the **First occupancy condition** apartments (9.32, 11.24, and 19.30 Euros) for **Berlin 2015**, **Berlin 2019**, and **Munich 2019** respectively, but it is at the highest with the **In need of renovation condition** apartments (14.80 Euros) for Munich 2015. Thus, rent_sqm is relatively higher with the first occupancy condition apartments in Berlin and Munich compared to other apartment condition categories.

Prediction of rent_sqm using four different scenarios

We will make predictions of **rent_sqm** in the main effect models given in Table 5.2, Table 5.3, Table 5.4, and Table 5.5 using four different scenarios: scenario 1 (smaller), scenario 2 (small), scenario 3 (large) and scenario 4 (larger) apartments. The goal is to investigate the behaviour of the **predicted rent_sqm** with respect to the different scenarios. We will use Table 5.15 and Table 5.16 to calculate the values of the quantitative variables in Table 5.17 which will be used for the prediction of **rent_sqm** in these different four scenarios. But for the categories of the qualitative variables, we will use the categories in Table 5.18 for our prediction based on the prediction result from Table 5.14. The Min, 25%, 50%, 75% and Max are respectively the minimum, first quartile (Q_1), median (Q_2), third quartile (Q_3) and maximum, which were defined in Chapter 3. The order of each summary table is the following: Berlin 2015, Berlin 2019, Munich 2015 and Munich 2019 respectively. In Table 5.16, we used the fact that the Midpoint of x and y is given by Midpoint $[x, y] = \frac{1}{2}(x + y)$. The predictions of these four scenarios are summarized in Table 5.19 and Table 5.20.

Table 5.15: Five-number summaries of the quantitative covariates

Variable	Summary				
addcost:	Min	25%	50%	75%	Max
	0.00	97.00	140.00	195.00	592.00
	0.00	100.00	141.00	200.00	599.00
	0.00	107.00	153.00	210.00	540.00
	0.00	120.00	170.00	220.00	550.00
heatcost					
	0.00	54.00	75.00	100.00	300.00
	0.00	49.00	65.00	90.00	300.00
	0.00	60.00	85.00	110.00	288.00
	0.00	55.00	80.00	109.00	300.00
conyear					
	1851	1910	1961	1992	2016
	1853	1918	1972	1998	2020
	1860	1962	1976	1999	2017
	1858	1965	1985	2014	2020
lmod					
	1983	2012	2014	2015	2016
	1982	2013	2016	2018	2018
	1981	2011	2014	2015	2016
	1983	2013	2015	2017	2018
lspace					
	23.00	55.00	69.00	89.00	161.00
	19.00	52.00	65.00	82.00	158.00
	23.00	55.00	71.00	90.00	161.00
	19.00	51.00	67.00	84.00	157.00
fspace					
	0.00	50.00	67.00	89.00	220.00
	0.00	48.00	65.00	87.00	250.00
	0.00	10.00	55.00	81.00	234.00
	0.00	11.00	55.00	82.00	249.00
energycon					
	0.00	88.00	117.00	149.00	350.00
	0.00	74.00	105.00	140.00	347.00
	0.00	85.00	122.00	155.00	338.00
	0.00	64.00	103.00	137.00	339.00
adlength					
	0.00	0.00	0.00	1.00	20.00
	0.00	0.00	0.00	1.00	20.00
	0.00	0.00	0.00	1.00	20.00
	0.00	0.00	0.00	1.00	20.00

Table 5.16: Mid points for the five-number summaries in Table 5.15

Variable	Summary			
addcost:	Midpoint [Min, Q_1]	Midpoint [Q_1, Q_2]	Midpoint [Q_2, Q_3]	Midpoint [Q_3, Max]
	48.50	118.50	167.50	393.50
	50.00	120.50	170.50	399.50
	53.50	130.00	181.50	375.00
	60.00	145.00	195.00	385.50
heatcost				
	27.00	64.50	87.50	200.00
	24.50	57.00	77.50	195.00
	30.00	72.50	97.50	199.00
	27.50	67.50	94.50	204.50
conyear				
	1880.50	1935.00	1976.50	2004.00
	1885.5	1945.00	1985.00	2009.00
	1911.00	1969.00	1987.50	2008.00
	1911.5	1975.00	199.50	2017.00
lmod				
	1997.50	2013.00	2014.50	2015.50
	1997.50	2014.50	2017.00	2018.00
	1996.00	2012.50	2014.50	2015.50
	1998.00	2014.00	2016.00	2017.50
lspace				
	39.00	62.00	79.00	125.00
	35.50	58.50	73.50	120.00
	39.0	63.0	80.5	125.5
	35.0	59.0	75.5	120.5
fspace				
	25.0	58.5	78.0	154.5
	24.0	56.5	76.0	168.5
	5.0	32.5	68.0	157.5
	5.5	33.0	68.5	165.5
energycon				
	44.0	102.5	133.0	249.5
	37.0	89.5	122.5	243.5
	42.5	103.5	138.5	246.5
	32.0	83.5	120.0	238.0
adlength				
	0.0	0.0	0.5	10.5
	0.0	0.0	0.5	10.5
	0.0	0.0	0.5	10.5
	0.0	0.0	0.5	10.5

Now, for **scenario 1**, we set all continuous variables to the average midpoints **([Min,** Q_1**])** in Table 5.16, for Berlin 2015, Berlin 2019, Munich 2015, and Munich 2019. For the addcost in scenario 1, we have $Senario\_1_{addcost} = \frac{1}{4}(48.50 + 50.00 + 53.50 + 60.00) = 53$

For **scenario 2**, we set all continuous variables to the average midpoints ($[Q_1$**,** $Q_2]$) in Table 5.16, for Berlin 2015, Berlin 2019, Munich 2015, and Munich 2019. For the addcost in scenario 2, we have $Senario\_2_{addcost} = \frac{1}{4}(118.50 + 120.50 + 130.00 + 145.00) = 129$

Also, for **scenario 3**, we set all continuous variables to the average midpoints ($[Q_2$**,** $Q_3]$) in Table 5.16, for Berlin 2015, Berlin 2019, Munich 2015, and Munich 2019. For the addcost in scenario 3, we have $Senario\_3_{addcost} = \frac{1}{4}(167.50 + 170.50 + 181.50 + 195.00) = 179$

Finally, for **scenario 4**, we set all continuous variables to the average midpoints ($[Q_3$**,** **Max**]) in Table 5.16, for Berlin 2015, Berlin 2019, Munich 2015, and Munich 2019. For the addcost in scenario 4, we have $Senario\_4_{addcost} = \frac{1}{4}(393.50 + 399.50 + 375.00 + 385.50) = 388$. These values are rounded up to their nearest whole numbers and are summarized in Table 5.17

Table 5.17: Values of quantitative variables used in predicting **rent_sqm** for the four scenarios in Table 5.19

Variables	scenario 1	scenario 2	scenario 3	scenario 4
addcost	53	129	179	388
heatcost	27	65	89	200
conyear	1897	1956	1987	2010
lmod	1997	2014	2016	2017
lspace	37	61	77	122
fspace	15	45	73	162
energycon	39	95	129	244
adlength	0	0	1	11

Table 5.18: Categories of the qualitative variables used in predicting **rent_sqm** for the four scenarios in Table 5.19

Variables	Categories	scenario 1	scenario 2	scenario 3	scenario 4
afloorg	(-1)-0		(-1)-0		
	1-2			1-2	
	3-9				3-9
	> 9	> 9			
	NA				
bfloorg	0-2		0-2		
	3	3			
	4				
	5				5
	>5			>5	
	NA				
nroomsg	1-1.5	1-1.5			
	2-2.5		2-2.5		
	3-3.5			3-3.5	
	> 3.5				> 3.5
nbedg	0-1	0-1	0-1		
	2			2	
	> 2				> 2
	NA				
nbathg	0-1	0-1	0-1		
	> 1			> 1	> 1
	NA				
elevatorg				Yes	Yes
	No	No	No		
	NA				
balconyg					
kitcheng	Yes		Yes	Yes	Yes
	No	No			
	NA				
ewwg	Yes			Yes	Yes
	No	No	No		
	NA				
subhg	Yes	Yes		Yes	
	No		No		No
	NA				
gtoiletg	Yes			Yes	Yes
	No	No	No		
gardeng	Yes	Yes		Yes	
	No		No		No
	NA				
hwwg	Yes			Yes	Yes
	No	No	No		
cellarg	Yes	Yes		Yes	
	No		No		No
parkingg	Yes			Yes	Yes
	No	No	No		
	NA				
furnishingg	Upscale			Upscale	Upscale
	Normal	Normal	Normal		
	NA				
eeffgg	High			High	High
	Meduim	Meduim			
	Low		Low		
	NA				
ecertg	consumption	consumption	consumption		
	building			building	building
petsg	Yes	Yes	Yes		
	No			No	No
	NA				
heatg	CH			CH	CH
	NCH	NCH	NCH		
	NA				
apcatg	top			top	
	middle		middle		
	low	low			
	below				below
	NA				
pcong	Md	Md			
	Mt		Mt		Mt
	First			First	
	Inr				
	NA				

Table 5.19: Prediction of **rent_sqm** using **four different scenarios** from Table 5.17 and Table 5.18 for **Berlin 2015**, **Berlin 2019**, **Munich 2015** and **Munich 2019 main effect models**

Scenarios	Berlin 2015	Berlin 2019	Munich 2015	Munich 2019
		increase in %		increase in %
scenario 1	4.71	5.12	14.68	14.95
%	100%	8.70 %	100%	1.84%
(90% confidence interval) (Lower limit, Upper limit)	(4.12 , 5.39)	(3.97 , 6.61)	(13.79 , 15.62)	(12.67 , 17.64)
scenario 2	6.42	8.87	13.49	14.74
%	100%	38.16 %	100%	9.27%
(90% confidence interval) (Lower limit, Upper limit)	(5.69 , 7.24)	(7.73 , 10.18)	(12.37 , 14.71)	(13.22 , 16.44)
scenario 3	7.27	9.39	13.11	22.38
%	100%	29.16 %	100%	70.71%
(90% confidence interval) (Lower limit, Upper limit)	(6.38 , 8.28)	(7.50 , 11.74)	(11.91 , 14.44)	(20.27 , 24.72)
scenario 4	12.61	21.37	11.94	19.12
%	100%	69.47 %	100%	60.13%
(90% confidence interval) (Lower limit, Upper limit)	(10.62 , 14.97)	(17.34 , 26.34)	(10.67 , 13.37)	(14.88 , 24.57)

We can now summarize the behaviour of the **predicted rent_sqm** in Table 5.19 as follows

- In **scenario 1**, the **predicted rent_sqm** increased in Berlin from 2015 to 2019 by 8.70% (108.70%) but increased in Munich by 1.84% (101.84%).
- In **scenario 2**, the **predicted rent_sqm** increased in Berlin from 2015 to 2019 by 38.16% (138.16%) but increased in Munich by 9.27% (109.27%).
- In **scenario 3**, the **predicted rent_sqm** increased in Berlin from 2015 to 2019 by 29.16% (129.16%) while in Munich, it increased by 70.71% (170.71%).
- In **scenario 4**, the **predicted rent_sqm** increased in Berlin from 2015 to 2019 by 69.47% (169.47%) but increased in Munich by 60.13% (160.13%).
- Also, in the **four scenarios**, the **predicted rent_sqm** in **Munich** is relatively higher than the **predicted rent_sqm** in **Berlin** except in the **fourth scenario** where the **predicted rent_sqm** in **Berlin** is higher than that of Munich
- Considering the prediction with the fourth scenario, from 2015 to 2019, we are 90% confident that rent_sqm increased by 69.47% and 60.13% in Berlin and Munich respectively with larger apartments.
- We equally noticed **an increase** in the **predicted rent_sqm** with respect to **time (2015 and 2019)** in **Berlin and Munich** for the four different scenarios.

Table 5.20: Prediction of **rent_sqm** using **four different scenarios** from Table 5.17 and Table 5.18 for **Berlin 2015**, **Berlin 2019**, **Munich 2015** and **Munich 2019 main effect models**

city/scenarios	**scenario 1**	**scenario 2**	**scenario 3**	**scenario 4**
		increase in % compared to scenario 1	**increase in % compared to scenario 1**	**increase in % compared to scenario 1**
Berlin 2015	4.71	6.42	7.27	12.61
%	100%	36.31 %	54.35 %	167.73%
(90% confidence interval) (Lower limit, Upper limit)	(4.12 , 5.39)	(5.69 , 7.24)	(6.38 , 8.28)	(10.62 , 14.97)
Berlin 2019	5.12	8.87	9.39	21.37
%	100%	73.24 %	83.40 %	317.38%
(90% confidence interval) (Lower limit, Upper limit)	(3.97 , 6.61)	(7.73 , 10.18)	(7.50 , 11.74)	(17.34 , 26.34)
Munich 2015	14.68	13.49	13.11	11.94
%	100%	-(100 - 91.89)%	-(100 - 89.31)%	-(100 - 81.34)%
(90% confidence interval) (Lower limit, Upper limit)	(13.79 , 15.62)	(12.37 , 14.71)	(11.91 , 14.44)	(10.67 , 13.37)
Munich 2019	14.95	14.74	22.38	19.12
%	100%	-(100 - 98.60)%	49.70 %	27.89%
(90% confidence interval) (Lower limit, Upper limit)	(12.67 , 17.64)	(13.22 , 16.44)	(20.27 , 24.72)	(14.88 , 24.57)

Table 5.20 is the transpose of Table 5.19 and we can also summarize the behaviour of the **predicted rent_sqm** in it with respect to the four scenarios as follows

- Looking at Berlin 2015 and 2019, we can see that the **predicted rent_sqm increased** from scenario 1 to scenario 2 by 36.31% and 73.24% (136.31% and 173.24%) respectively but in Munich 2015 and 2019, the **predicted rent_sqm decreased** by 8.11% and 1.4% (91.89% and 98.60%) respectively from scenario 1 to scenario 2.
- In Berlin 2015 and 2019, we can also see that the **predicted rent_sqm increased** from scenario 1 to scenario 3 by 54.35% and 83.40% (154.35% and 183.40%) respectively while in Munich 2015 and 2019, the **predicted rent_sqm decreased** by 10.69% but **increased** by 49.70% (89.31% and 149.70%) respectively from scenario 1 to scenario 3.
- We can also notice that from scenario 1 to scenario 4, the **predicted rent_sqm increased** by 167.73% and 317.38% (267.73% and 417.38%) in Berlin 2015 and 2019 respectively while

in Munich 2015, the **predicted rent_sqm decreased** by 10.69% (81.34%) but **increased** by 27.89% (127.89%) in Munich 2019.

- Looking at Berlin 2015 and Munich 2015, we see that in Berlin 2015, the **predicted rent_sqm increased** from scenario 1 to scenario 2, scenario 1 to scenario 3, and scenario 1 to scenario 4 by 36.31%, 54.35% and 67.73% (136.31% 154.35% and 267.73%) respectively while in Munich 2015, **predicted rent_sqm decreased** by 8.11%, 10.69% and 18.66% (91.89%, 89.31%, and 81.34%) respectively.

- Also, looking at Berlin 2019 and Munich 2019, we see that in Berlin 2019, the **predicted rent_sqm increased** from scenario 1 to scenario 2, scenario 1 to scenario 3, and scenario 1 to scenario 4 by 73.24%, 83.40% and 317.38% (173.24%, 183.40% and 417.38%) respectively while in Munich 2019, **predicted rent_sqm decreased** by 1.40%, increased by 49.70% and 27.89% (98.60%, 149.70%, and 127.89%) respectively.

- Thus, there is an **an increasing trend** in the **predicted rent_sqm** with respect to the four scenarios for **Berlin 2015 and 2019**. On the other hand, there is **a decreasing trend** in the **predicted rent_sqm** for the four scenarios in **Munich 2015** but a **fluctuation in the predicted rent** in the four scenarios for **Munich 2019**.

- We are 90% confident that the **predicted rent_sqm** increased from scenario 1 to scenario 3 by 54.35% and 83.40% in Berlin 2015 and 2019 respectively but increased from scenario 1 to scenario 4 by 167.73% and 317.38% in Berlin 2015 and 2019 respectively unlike in Munich 2015, where the **predicted rent_sqm** decreased by 18.66% from scenario 1 to scenario 4 but it increased by 49.70% from scenario 1 to scenario 3 in Munich 2019.

Chapter 6

Summary of findings

In this chapter, we will summarize the result findings in the modeling of **rent_sqm** for Berlin and Munich rental properties in 2015 and 2019.

6.1 Summary of findings

Considering the results from this thesis, we summarize our findings as follows

- Munich has more apartments with a balcony, a kitchen, a higher number of bathrooms, a higher number of bedrooms, a garden, the inclusion of warm water consumption in the heating cost value calculation, a parking space, upscale and normal furnishing, maintained condition, and do not allow pets compared to Berlin. On the other hand, Berlin has more apartments with guest toilets, the inclusion of warm water consumption in the energy consumption value calculation, certificate of eligibility to public housing, and building type of energy performance certificate than Munich
- The price of **rent_sqm** of an apartment relatively increases in both Berlin and Munich with apartments that have any of the following features; 5 building floors, a parking space, an elevator, a kitchen, Upscale furnishing, central heating type, no pets allowed, guest toilet, below apartment category and First occupancy condition (vice versa). It equally increases with respect to time in both cities. For instance, the **predicted rent_sqm** increased in Berlin from 2015 to 2019 by 29.92%, 24.26%, 29.22%, 15.94%, 24.26%, 24.26%, 13.81%, and 20.60% with apartments that have 5 building floors, 2-2.5 rooms, an elevator, a kitchen, Upscale furnishing, central heating type, below apartment category and First occupancy condition respectively, while in Munich it also increased from 2015 to 2019 by 31.17%, 31.17%, and 39.86% with apartments that have a kitchen, Upscale furnishing, and First occupancy condition.
- Also, **rent_sqm** price relatively increases in Berlin with respect to the order of the categories of the following features; the number of bedrooms (0-1, 2,>2), the inclusion of warm water consumption in the energy consumption value calculation (No, Yes), the energy efficiency

rating (Low, Medium, High), the number of bathrooms (0-1, >1), and a garden (No, Yes) (vice versa). On the other hand, it relatively decreases in Munich with respect to the order of the categories of the following features; the number of bedrooms (0-1, 2,>2), the inclusion of warm water consumption in the energy consumption value calculation (No, Yes), the energy efficiency rating (Low, Medium, High), the energy performance certificate (building, consumption), and a parking space (Yes, No) (vice versa).

- Considering the predictions with our four scenarios, scenario 1, scenario 2, scenario 3 and scenario 4 (smaller, small, large, and larger apartments), the **predicted rent_sqm** increased from 2015 to 2019 in Berlin by 8.70%, 38.16%, 29.16%, and 69.47% for smaller, small, large, and larger apartments respectively but in Munich it increased in the same period of time by 1.84%, 9.27%, 70.71%, and 60.13% for smaller, small, large, and larger apartments respectively.

- Finally but not the least, in Berlin 2015, the **predicted rent_sqm increased** from scenario 1 to scenario 2, scenario 1 to scenario 3, and scenario 1 to scenario 4 by 36.31%, 54.35% and 67.73% respectively while in Munich 2015, the **predicted rent_sqm decreased** by 8.11%, 10.69% and 18.66% respectively. On the other side, in Berlin 2019, the **predicted rent_sqm increased** from scenario 1 to scenario 2, scenario 1 to scenario 3, and scenario 1 to scenario 4 by 73.24%, 83.40% and 317.38% respectively while in Munich 2019, it **decreased** by 1.40%, increased by 49.70% and 27.89% respectively.

6.2 Conclusion

The price of **rent_sqm** increases with respect to time in both cities and is relatively higher in Munich compared to Berlin. In Berlin, rent increases at an increasing rate but increases at a decreasing rate in Munich with respect to bigger apartments. Upscale furnishing is another major factor that equally increases the price of **rent_sqm** based on its positive significance in all our fitted models.

6.3 Recommendations

We would like to make the following recommendations based on our findings

- We recommend that families looking for bigger apartments can reside in Munich as they will pay less compared to residing in Berlin (vice versa).

- People wishing to have more savings from their rent budget, can go for the Normal furnishing apartments instead of the Upscale furnishing apartments.

- It would be nice if many smaller apartments are built in Munich instead of bigger apartments to accommodate the huge number of people coming into Munich. It can help to reduce this increase in **rent_sqm** with smaller apartments in Munich.

Appendix A

Additional EDA plots

Figure A.1: Histograms of response variable - **rent_sqm**: first column = **counts**, second column = **percentage**

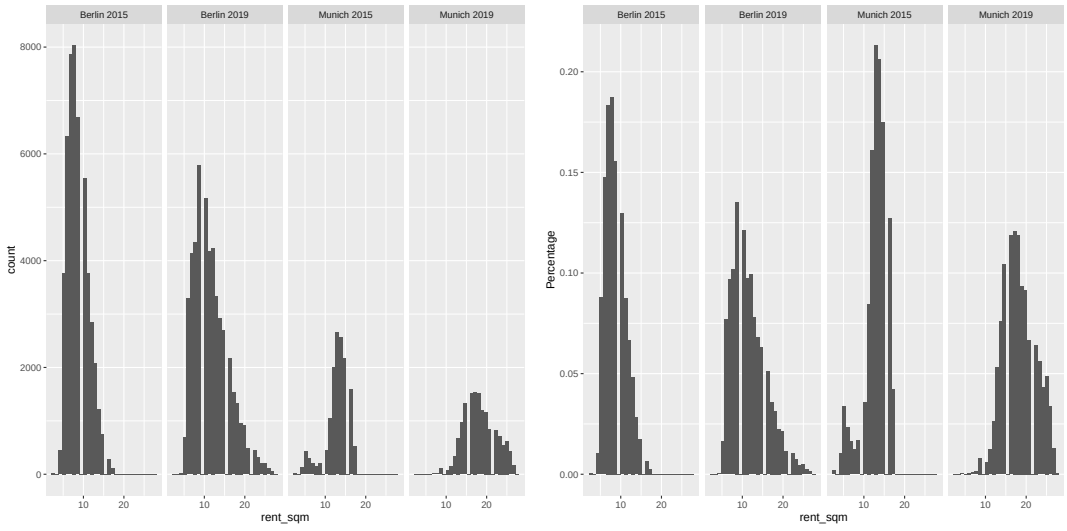


Figure A.2: Histograms of quantitative covariates (part I): first column = **counts**, second column = **percentage**

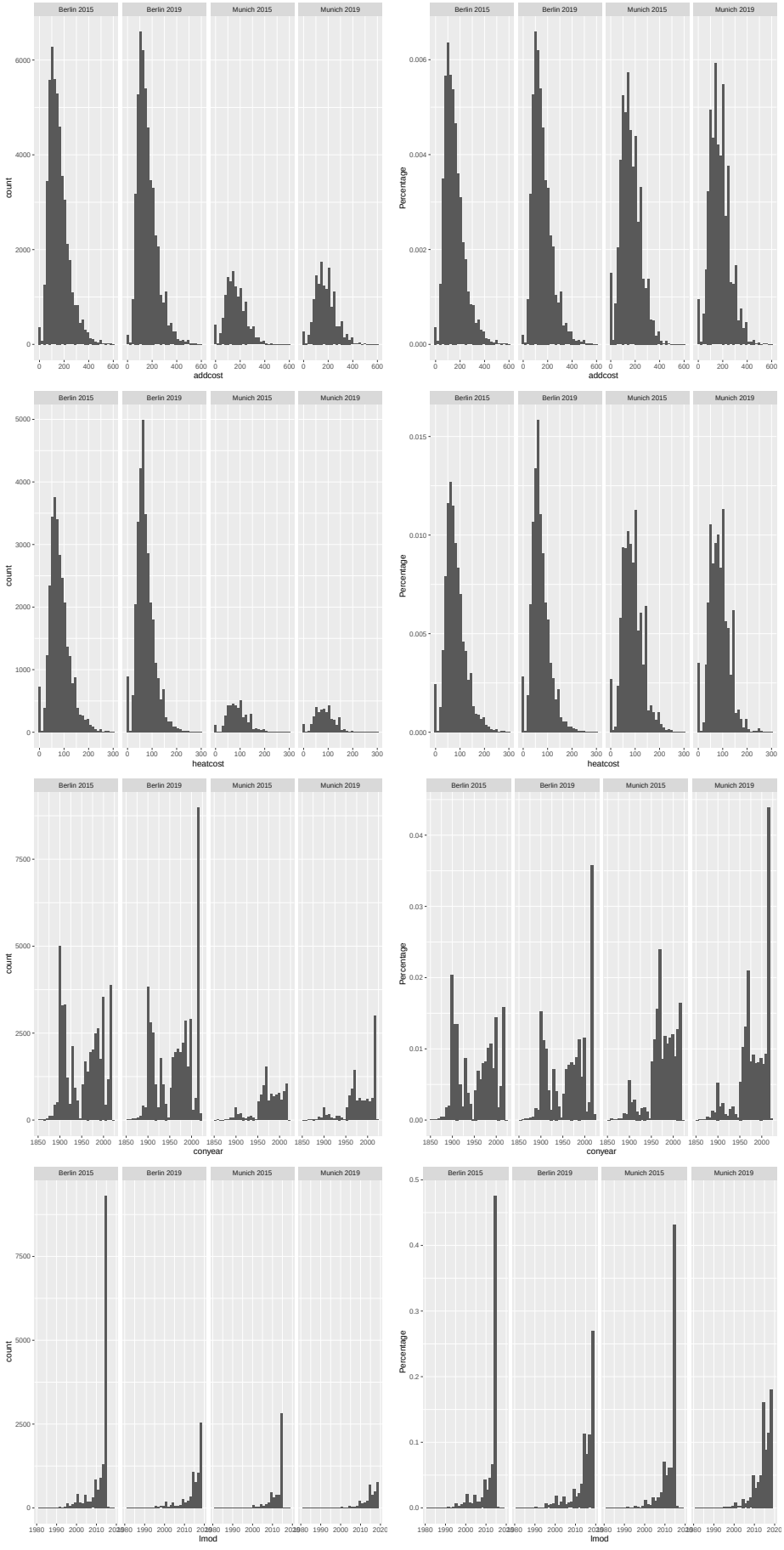


Figure A.3: Histograms of quantitative covariates (part II): first column = **counts** second column = **percentage**

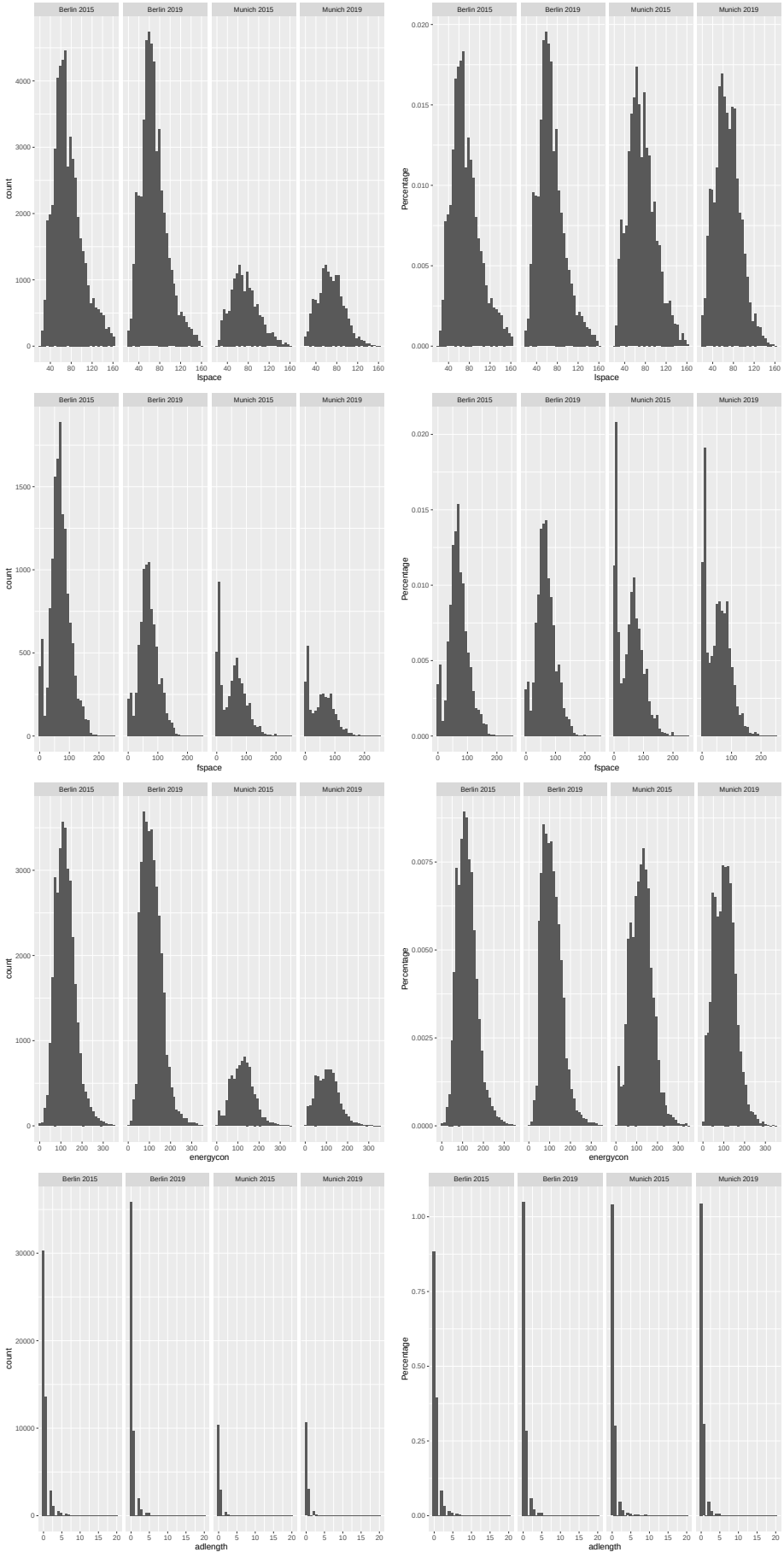


Figure A.4: Bar plot of qualitative variables(part II)

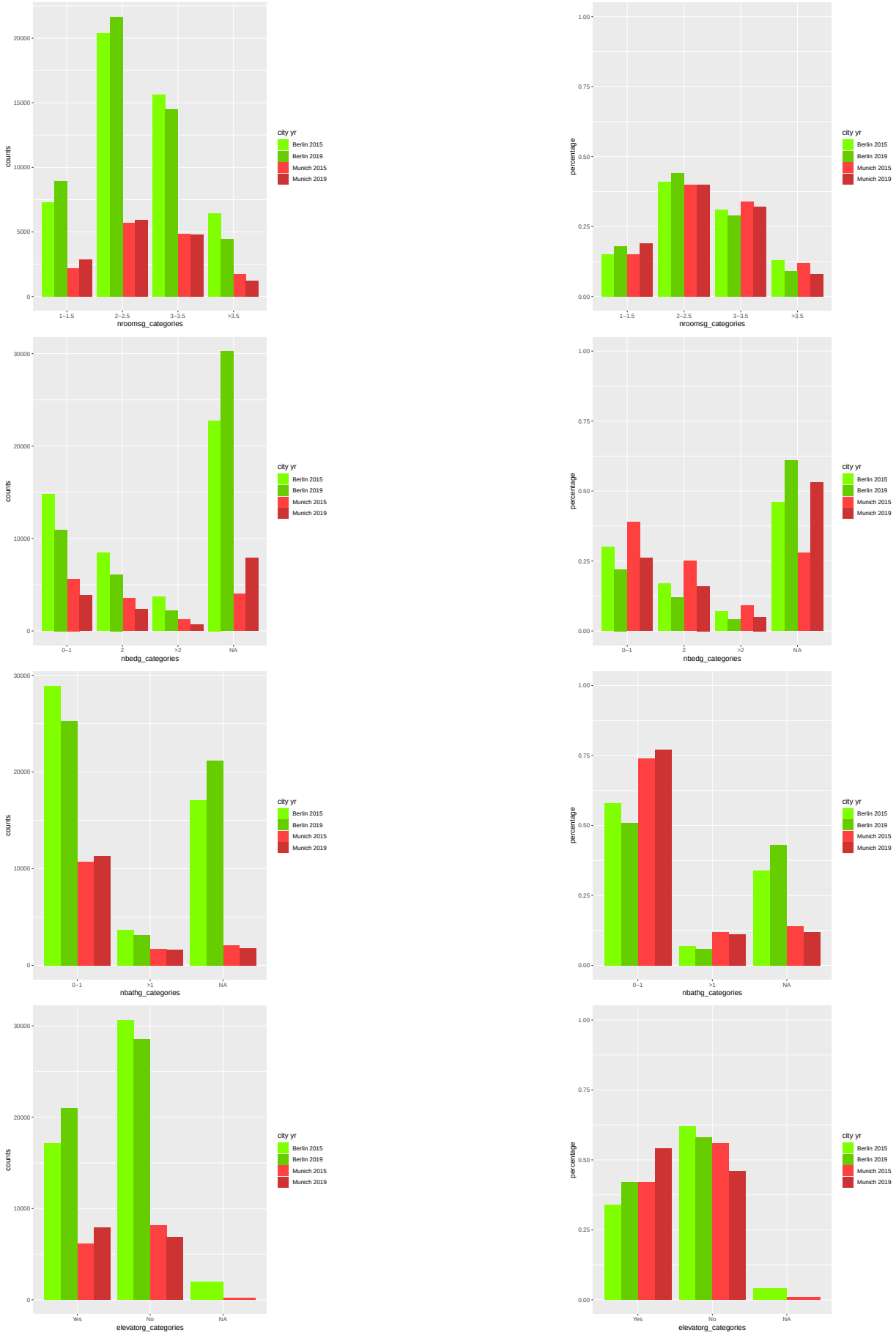


Figure A.5: Bar plot of qualitative variables (part III)

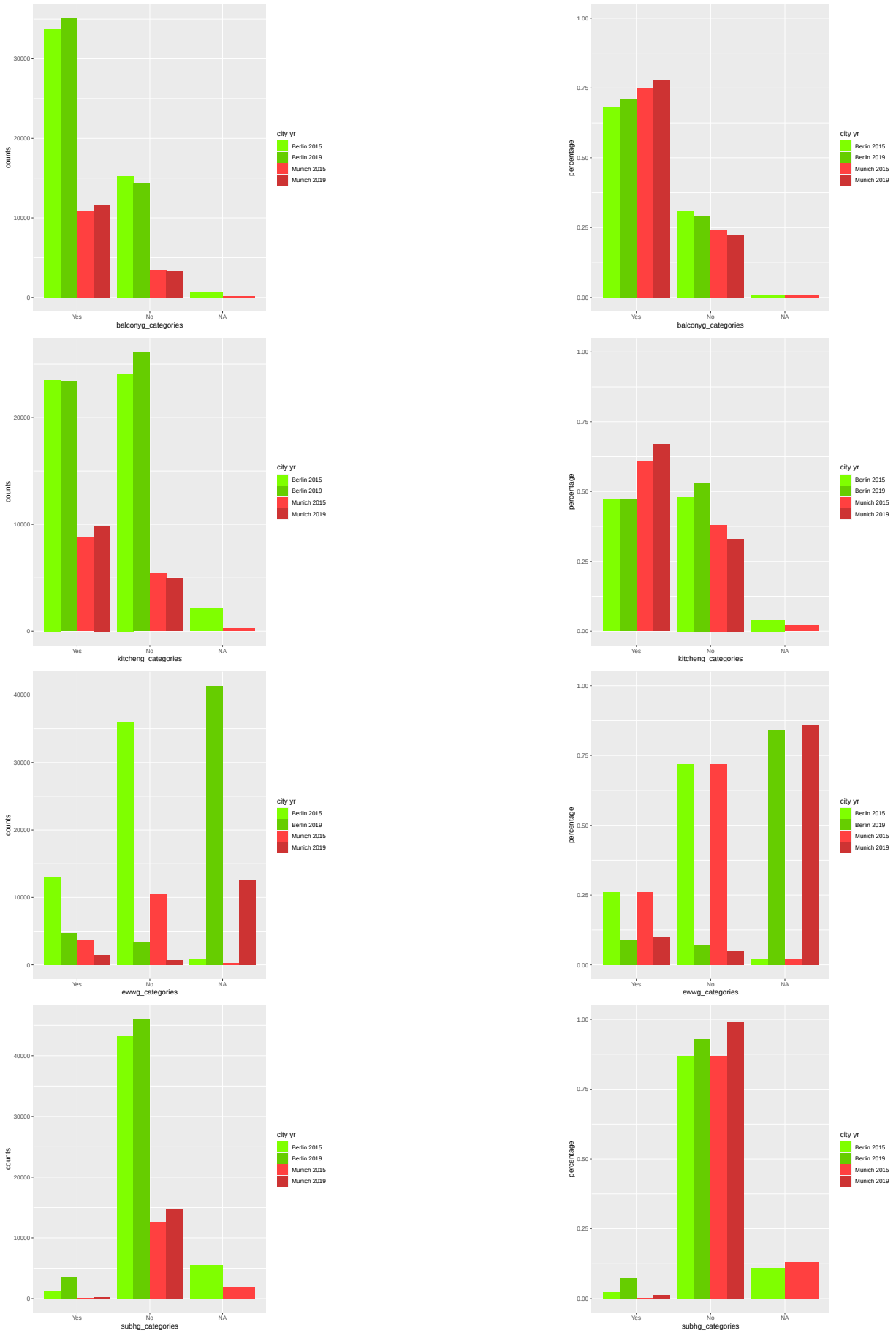


Figure A.6: Bar plot of qualitative variables (part IV)

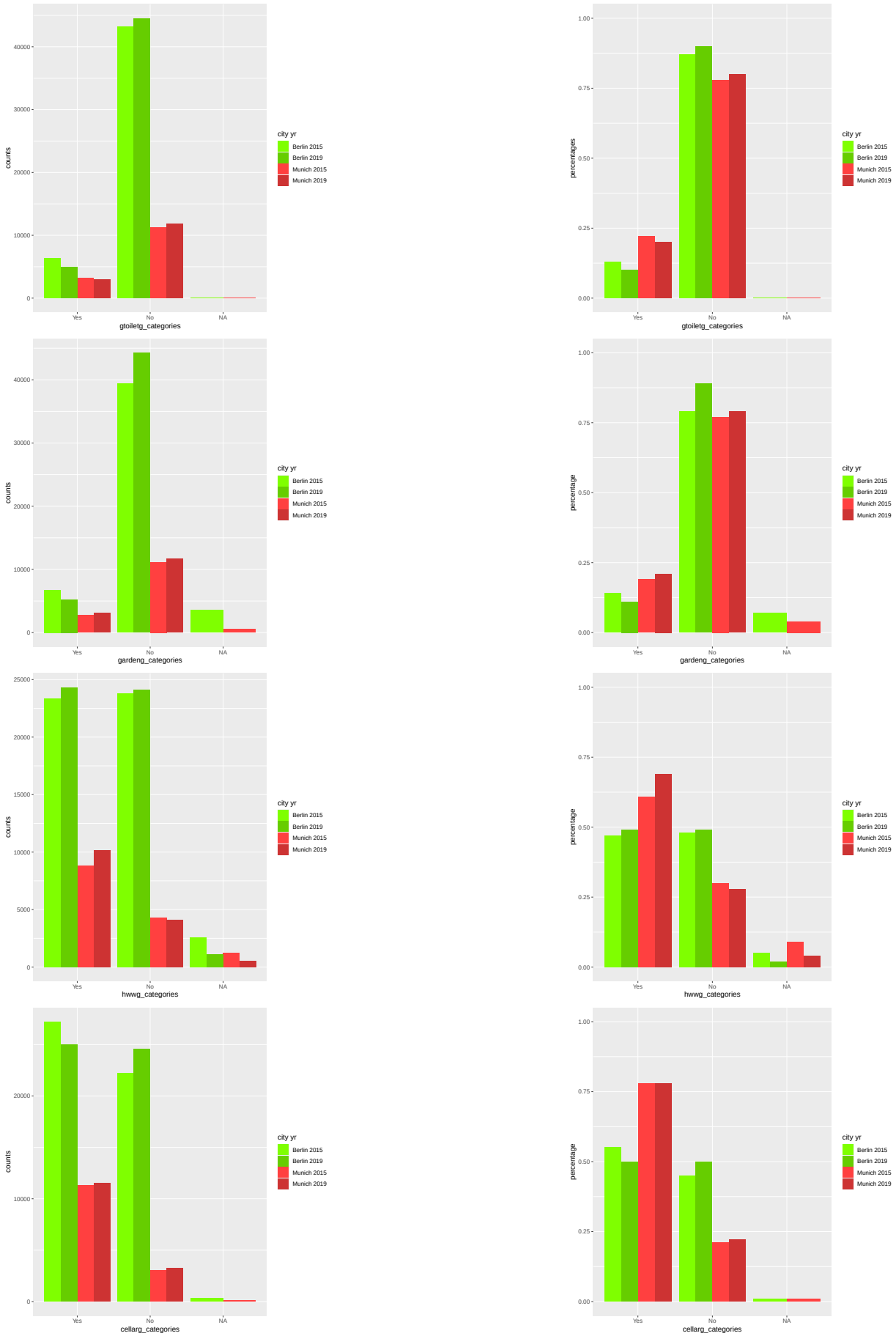


Figure A.7: Bar plot of qualitative variables (part V)

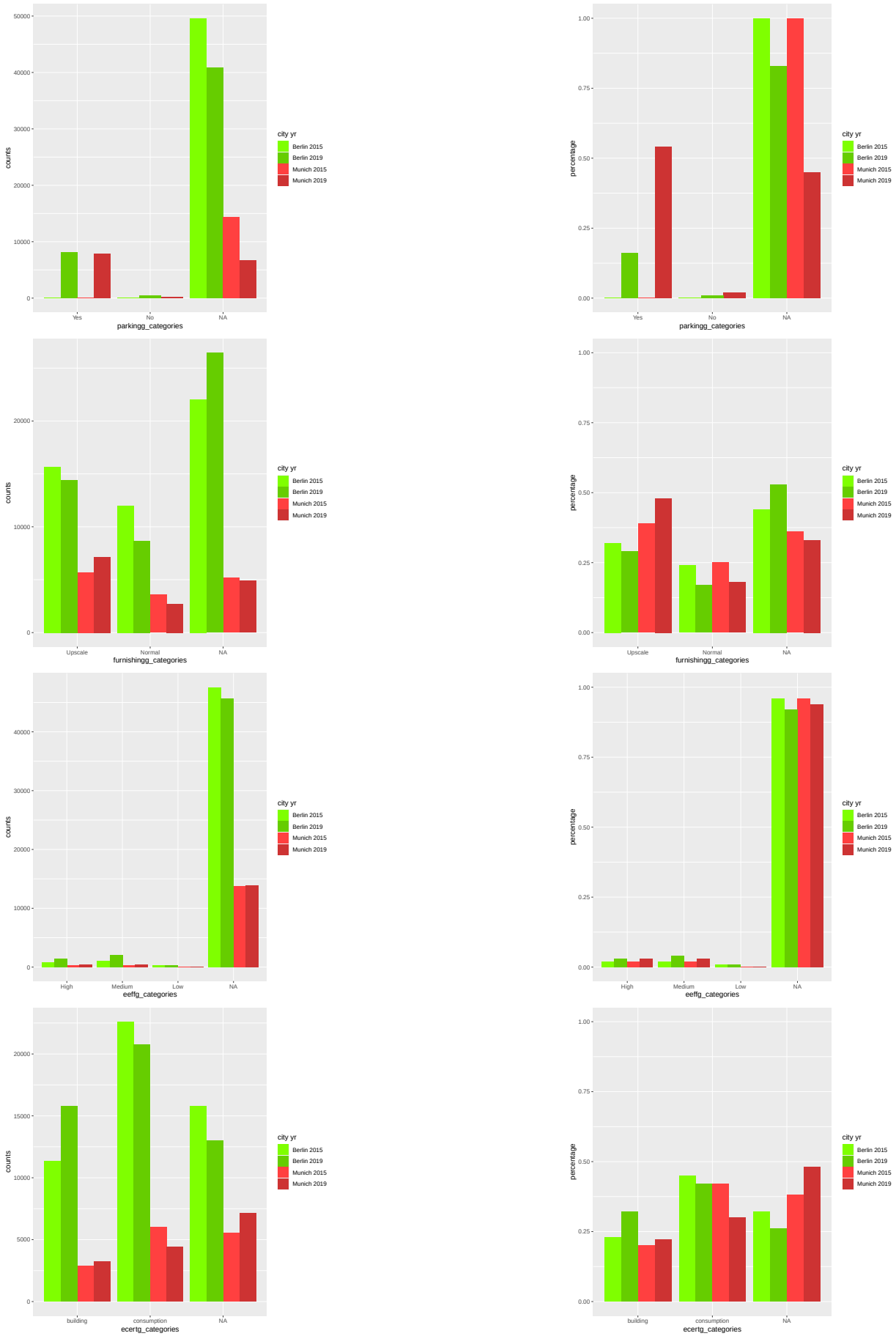


Figure A.8: Bar plot of qualitative variables (part VI)

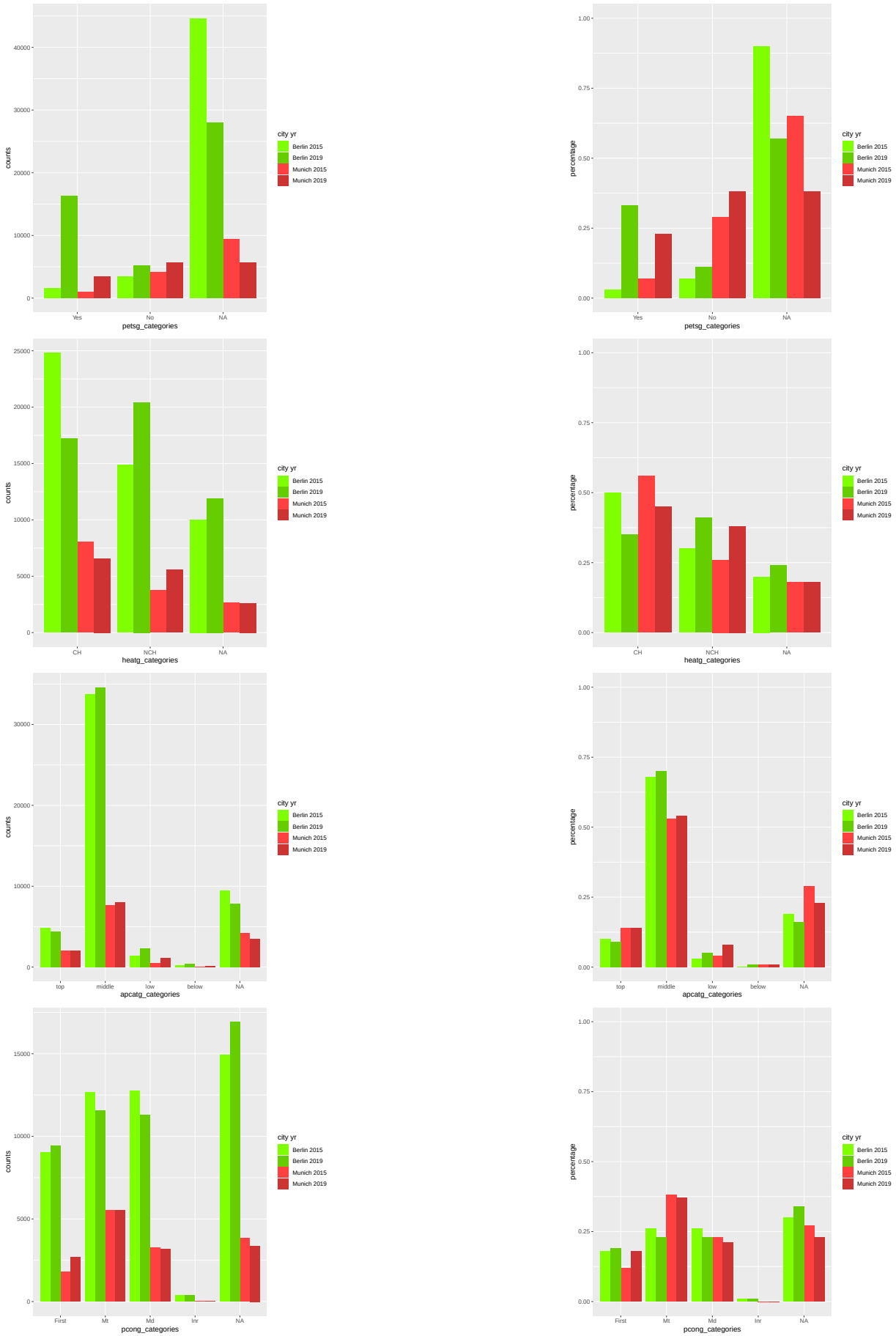


Figure A.9: Scatter plots of quantitative covariates versus response (rent_sqm) with **linear smooth**: first column = **(rent_sqm)** and second column = **log(rent_sqm)**. (first row) = **Berlin 2015**, (second row)= **Berlin 2019**, (third row) = **Munich 2015**, (fourth row) = **Munich 2019**

Figure A.10: Scatter plots of quantitative covariates versus response (rent_sqm) with **linear smooth**: first column = **log(log(rent_sqm))** and second column = **1/(rent_sqm)**. (first row) = **Berlin 2015**, (second row)= **Berlin 2019**, (third row) = **Munich 2015**, (fourth row) = **Munich 2019**

List of Figures

List of Tables

Bibliography

Abraham, B. and Ledolter, J. (2006). *Student solutions manual for Introduction to regression modeling*. University of Iowa.

Allen, M. P. (2004). *Understanding regression analysis*. Springer Science & Business Media.

Brown, J. D. (2014). *Linear models in matrix form*. Springer.

Christensen, R. (1996). *Analysis of variance, design, and regression: applied statistical methods*. CRC Press.

Christensen, R. (2018). *Analysis of variance, design, and regression: Linear modeling for unbalanced data*. Chapman and Hall/CRC.

Czado and Brechmann (2021). *Lecture Slides on GLM, Study material from the research group Mathematical Statistics in the Department of Mathematics at the Technical University Munich Deutschland*. `https://www.groups.ma.tum.de/statistics/personen/claudia-czado/forschung/lecture-slides/`.

Fahrmeir, L., Kneib, T., Lang, S., and Marx, B. (2013). Regression models. In *Regression*, pages 21–72. Springer.

Gustafsson, A. and Wogenius, S. (2014). Modelling apartment prices with the multiple linear regression model.

Horton, N. J. and Kleinman, K. (2015). *Using R and RStudio for data management, statistical analysis, and graphics*. CRC Press.

Kholodilin, K. A. and Mense, A. (2012). German cities to see further rises in housing prices and rents in 2013. *DIW Economic Bulletin*, 2(12):16–26.

Lin, D., Wei, L., and Ying, Z. (2002). Model-checking techniques based on cumulative residuals. *Biometrics*, 58(1):1–12.

Lindsey, J. K. (2000). *Applying generalized linear models*. Springer Science & Business Media.

Lutz, E. (2020). The housing crisis as a problem of intergenerational justice: The case of germany.

McNeil, K. A., Newman, I., and Kelly, F. J. (1996). *Testing research hypotheses with the general linear model*. SIU Press.

Mobert, J. (2017). Outlook of the german housing market in 2017. *Outlook*.

Möbert, J., Schneider, S., and AG, D. B. (2018). The german housing market in 2018. *Deutsche Bank Research*.

Nelder, J. A. and Wedderburn, R. W. (1972). Generalized linear models. *Journal of the Royal Statistical Society: Series A (General)*, 135(3):370–384.

Osborne, J. W. and Waters, E. (2002). Four assumptions of multiple regression that researchers should always test. *Practical assessment, research, and evaluation*, 8(1):2.

Ricci, L. (2010). Adjusted r-squared type measure for exponential dispersion models. *Statistics & probability letters*, 80(17-18):1365–1368.

Seber, G. A. (2015). *The Linear Model and Hypothesis*. Springer.

Thomschke, L. (2015). Changes in the distribution of rental prices in berlin. *Regional Science and Urban Economics*, 51:88–100.

Vik, P. (2013). *Regression, ANOVA, and the general linear model: A statistics primer*. SAGE Publications.