TU Dortmund

Introductory Case Studies

Project 3: Regression Modelling

Lecturers:
Prof. Dr. Jörg Rahnenführer
JProf. Dr. Antonia Arsova
Dr. Birte Hellwig
M.Sc. Julia Duda
M.Sc. Sven Pappert

Author: Hassan Ali

Group number: 1

Group members: Heba Alhosainy, Hendrik Linn, Hassan Ali

February 5, 2021

Contents

1. Introduction

Several cities across Germany have their own ‘Mietspiegel’ or ‘rent index’. These indices provide tenants and landlords an overview of the housing market situation in a given area. A noteworthy example is that of Munich; which for the past 20 years, has been one of the most expensive cities for renters in the country (Kristie 2019). Munich’s official city portal outlines its rent index (Wohnen und Migration München 2019). The website also features an online calculation program for the total rent of a property by prompting to insert its details such as its size, year of construction, the building type, etc. This estimation of the average net rent of a property from its characteristics constitutes a typical regression problem.

This report is closely related to the above-mentioned example of rental price evaluation. However, instead of focusing on just one particular city, it deals with the rental price data for one whole state of Germany. Here, the dataset comprises of rental offers from the entire province of North Rhine-Westphalia. Furthermore, rather than evaluating the total rent of a property, a linear model for estimating the rental price per square meter of the properties is built. Besides this, a logistic model is also trained to predict whether a property is newly constructed (i.e. constructed in the year 2019 or 2020) or not.

The statistical analysis of the data comprises of four parts. Firstly, the data is preprocessed to deal with missing values in some of the variables and to transform some of them by grouping their outcomes into new categories. In the second part, the univariate analysis of the variables is carried out to investigate the dataset’s underlying structure. This is followed by analysing the trained linear model for the rental price per square meter, and its assumptions, in part three. Finally in part four, the logistic model which predicts whether a given rental offer was built recently or not is evaluated.

The goal of the report is to train and assess the aforementioned linear and logistic regression models for rent per square meter and the binary variable ‘newly constructed’. Simple regression models are trained without employing any polynomial effects of the covariates or their interactions. In both cases, a best possible model is estimated using backward stepwise variable selection while employing the Akaike Information Criterion as a model selection criterion. The coefficients of the trained models are then interpreted to show how the covariates affect the response variables. Lastly, the predictive power of the linear model is judged by the adjusted R-squared coefficient, while that of the logistic model is deduced by generating a confusion matrix.

The resulting linear model is able to explain only about 37% of the variance in the rent per square meter of the properties. On the other hand, the logistic model has a discriminatory power of more than 94%. However, it is only able to classify around 44% of the newly constructed flats correctly. This could be explained by the fact that there are comparatively few training examples for this category; less than 7% of all the rental offers in the dataset comprise of newly constructed properties.

Besides this Introduction, the report consists of four more sections. The Problem Statement section describes the variables in the dataset in detail and discusses the data quality. For instance, around 19% of the rental offers have missing values for the total rent, the variable used to compute the response variable for the linear model. The Methods portion of the report provides formulas and assumptions for the linear and logistic regression models. It also outlines the process of backward selection, the Akaike Information Criterion, the coefficient of determination and the confusion matrix. The Statistical Analysis section describes the preprocessing of the dataset, provides univariate analyses of all variables and interprets the coefficients of the final linear and logistic regression models. Finally, the Summary portion concisely rehearses the most important results. It discusses potential improvements to the experiment, such as collecting more data, and suggests ways in which it can be extended, e.g. by also considering interactions between the covariates in the regression models.

2. Problem Statement

The data is collected from the German real estate web-portal Immobilienscout24. The website features rental offers as well as homes for sale. The dataset used in this report comprises of 12118 rental offers as of 20 February 2020. All properties are located in the province of North Rhine-Westphalia, Germany. The full dataset is available on `https://www.kaggle.com/`.

The dataset consists of 16 variables; ten of these being categorical and six numeric. There are five binary categorical variables which assume the values true or false. When the variable ‘newlyConst’ is true, this means that a property is newly constructed (i.e. constructed in 2019 or 2020), and vice versa. Similarly, when the variables ‘balcony’, ‘hasKitchen’, ‘lift’ or ‘garden’ are set to true, this implies that a property has a balcony, a kitchen, a lift or a garden; and vice versa.

Three other nominal variables also have two levels. The variable ‘condition’ which indicates the condition in which the property is, takes on the values average and good. The variable ‘lastRefurbish’ specifies the time period in which a given property was last renovated, with levels last 5 years and over 5 years ago. The variable ‘energyEfficiencyClass’ indicates the energy efficiency class of the building and has the levels A+/A/B/C and D/E/F/G/H.

The two remaining categorical variables have more than 2 levels. The variable ‘typeOfFlat’ indicates the type of flat and has 10 levels: roof storey, apartment, ground floor, half basement, penthouse, terraced flat, maisonette, raised ground floor, loft and other. The other variable ‘regio2’ refers to the city in which a property is located and has 54 levels.

Out of the six numeric variables, four are discrete. The variable ‘ID’ is a unique arbitrary identification number assigned to each rental offer. It assumes the values of the natural numbers from 1 to 12118. The variable ‘yearConstructed’ indicates the year of construction of a property. The variable ‘noParkSpaces’ refers to the number of parking spaces provided with a rental offer. The variable ‘floor’ indicates the floor in which a property is located. For this variable, a value of -1 means that the property is located in a basement; a 0 indicates ground floor; a 1 indicates the first floor and so on.

The remaining two numeric variables are continuous. The variable ‘totalRent’ refers to the total rent of a property, which includes its base rent, service charges and heating costs. And finally, the variable ‘livingSpace’ indicates the size of a property in square meters.

The data quality suggests that preprocessing is required before training the linear or logistic regression models. For instance, many of the variables in the dataset have missing values. For ‘totalRent’ these amount to 2284; for ‘noParkSpaces’ 7931; for ‘typeOfFlat’ 714; for ‘lastRefurbish’ 8088; for ‘condition’ 2845; and for ‘energyEfficiencyClass’ 8457 in total. Hence, during the preprocessing step, these missing values have to be treated in a meaningful way.

The objectives of the report are, first of all, to preprocess the data and carry out a univariate analysis of the variables. Thereafter, a linear model is built for the rent per square meter of the properties as the response variable, and a logistic model is trained to predict whether a property is newly constructed or not. In both cases, the predictive power of the models is judged, their shortcomings are highlighted, and possible areas of improvement are identified.

3. Statistical Methods

The following statistical methods and mathematical formulas are used. The statistical software R (R Development Core Team 2020), version 4.0.3 has been used for analysis.

3.1. Multiple Linear Regression Model

Multiple linear regression models the effect of a vector of k independent variables or covariates, $x_1, ..., x_k$, on a dependent or response variable y. Here, the response variable is continuous while the covariates can be either continuous or appropriately coded categorical variables. The response variable is not a deterministic function $f(x_1, ..., x_k)$ of the covariates. Instead, this relationship shows random errors.

$$y = f(x_1, ..., x_k) + \varepsilon_i = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k + \varepsilon_i$$

The linear function f is called the systematic component of the model while the error term ε is referred to as the random or stochastic component. To model a categorical covariate $x \in \{1, ..., c\}$ with c categories, category c can be treated as a reference. Thereafter, $c - 1$ dummy variables can be defined as follows and then included in the model.

$$x_{i1} = \begin{cases} 1 & x_i = 1 \\ 0 & \text{otherwise} \end{cases} \quad \dots \quad x_{i,c-1} = \begin{cases} 1 & x_i = c-1 \\ 0 & \text{otherwise} \end{cases}$$

The parameters $\beta_0, ..., \beta_k$ are unknown. By combining the covariates and the parameters into separate $p = k + 1$ dimensional vectors, $\boldsymbol{x} = (1, x_1, ..., x_k)'$ and $\boldsymbol{\beta} = (\beta_0, ..., \beta_k)'$, the systematic component can be expressed as a vector product. Therefore,

$$y = f(\boldsymbol{x}) + \varepsilon = \boldsymbol{x'\beta} + \varepsilon$$

To estimate the parameters, data $(y_i, x_{i1}, ..., x_{ik})$ is collected where $i = 1, ...n$. The vectors $\boldsymbol{y}$ and $\boldsymbol{\varepsilon}$ and the design matrix $\boldsymbol{X}$ are defined as follows:

$$\boldsymbol{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}, \quad \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

$$\boldsymbol{X} = \begin{pmatrix} 1 & x_{11} & \dots & x_{1k} \\ \vdots & \vdots & & \vdots \\ 1 & x_{n1} & \dots & x_{nk} \end{pmatrix}$$

Here the errors are normally distributed, with zero mean and a constant variance across them (i.e. homoscedastic errors), $\boldsymbol{\varepsilon} \sim (\mathbf{0}, \sigma^2 \boldsymbol{I})$. The design matrix $\boldsymbol{X}$ is assumed to have full column rank, implying that all columns are linearly independent. Hence, n equations can be formed, as follows

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}$$

The unknown parameters $\boldsymbol{\beta}$ are estimated using the method of least squares. Here the estimates $\hat{\boldsymbol{\beta}}$ are the minimizers of the sum of squared deviations.

$$\text{LS}(\boldsymbol{\beta}) = (\boldsymbol{y} - \boldsymbol{X\beta})'(\boldsymbol{y} - \boldsymbol{X\beta})$$

Setting the derivative of the above expression with respect to $\boldsymbol{\beta}$ equal to zero, leads to the unique solution of the least squares estimator.

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X'X})^{-1}\boldsymbol{X'y}$$

(Fahrmeir et al. 2013, p. 74-107).

To test for the significance of a parameter β_j, the hypotheses are $H_0 : \beta_j = 0$ against $H_1 : \beta_j \neq 0$. The estimated t-statistic is calculated as follows:

$$\hat{t}_j = \frac{\hat{\beta}_j}{\hat{se}_j},$$

where $\hat{se}_j = [\widehat{\text{Var}(\hat{\beta}_j)}]^{1/2}$ is the estimated standard error of $\hat{\beta}_j$. For a predefined significance level of α (in this report 0.05), the absolute value of the above statistic is compared to the $(1 - \alpha/2)$th quantile of the t-distribution with $n - p$ degrees of freedom. H_0 is rejected if:

$$|\hat{t}_j| > t_{1-\alpha/2}(n - p)$$

(Fahrmeir et al. 2013, p. 135).

3.2. Coefficient of Determination

For a linear regression model, the coefficient of determination R^2 is defined as

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})}{\sum_{i=1}^{n}(y_i - \bar{y})} = 1 - \frac{\sum_{i=1}^{n} \hat{\varepsilon}^2}{\sum_{i=1}^{n}(y_i - \bar{y})}$$

where $\bar{y}$ is the mean value of the response variable and $\hat{y}_i$ for $i = 1, ..., n$ are its estimated values (Fahrmeir et al. 2013, p. 115).

R^2 has the range, $0 \leq R^2 \leq 1$. When R^2 is closer to 1, the residual sum of squares is smaller and the fit to the data is better. If R^2 is closer to 0, this sum is larger and the regression model is poorly fitted. R^2 is a measure of the proportion of the variance in the response variable that is predictable from the covariates.

The coefficient of determination is of limited value when it comes to model comparison. The corrected coefficient of determination mitigates its shortcomings by incorporating a correction term in the formula to account for the number of parameters in the model.

$$\bar{R}^2 = 1 - \frac{n-1}{n-p}(1 - R^2)$$

(Fahrmeir et al. 2013, p. 147-148).

3.3. Logistic Regression (Logit) Model

As for the linear regression model, it is assumed that the data consists of the form $(y_i, x_{i1}, ..., x_{ik})$ where $i = 1, ...n$. But here the response variable y is binary coded with true and false or 1 and 0. The goal is to model the probability of the outcome $y_i = 1$ conditioned on the values of the covariates $x_{i1}, ..., x_{ik}$:

$$E(y_i) = \pi_i = P(y_i = 1 | x_{i1}, ..., x_{ik}) = \frac{\exp(\eta_i)}{1 + \exp(\eta_i)}$$

where η_i is the linear predictor given by

$$\eta_i = \beta_0 + \beta_1 x_{i1} + ... + \beta_k x_{ik} = \boldsymbol{x}_i' \boldsymbol{\beta}$$

The logarithmic odds of $y = 1$ against $y = 0$ is a linear model given by,

$$log(\frac{\pi}{1-\pi}) = \eta = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k$$

Taking the exponential of both sides of the equation leads to,

$$\frac{\pi}{1-\pi} = \exp(\beta_0)\exp(\beta_1 x_1)...\exp(\beta_k x_k)$$

This implies that the covariates affect the odds $\pi/(1-\pi)$ in an exponential-multiplicative fashion. If, for example, a covariate x_j increases by 1, then the odds of $y = 1$ against $y = 0$ are now $\exp(\beta_j)$ times the original odds (Fahrmeir et al. 2013, p. 270-271).

The logit model does not fulfill the assumptions of the model for multiple linear regression. Therefore, instead of the least squares estimator, the maximum likelihood estimator is used. This maximizes the likelihood function given by

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} {\pi_i}^{y_i} (1-\pi_i)^{1-y_i}$$

Taking the logarithm of the likelihood function yields the log-likelihood

$$l(\boldsymbol{\beta}) = \sum_{i=1}^{n} l_i(\boldsymbol{\beta}) = \sum_{i=1}^{n} \{y_i \text{log}(\pi_i) - y_i \text{log}(1-\pi_i) + \text{log}(1-\pi_i)\}$$

To find out the maximum likelihood estimator, the partial derivatives of the log-likelihood are evaluated and set to zero. This leads to a p-dimensional system of non-linear equations for the estimates $\hat{\boldsymbol{\beta}}$. Popular iterative methods for solving these equations include the Newton-Raphson and the Fisher scoring algorithm (Fahrmeir et al. 2013, p. 280-281).

In logistic regression, the response variable can be said to follow a binomial distribution, $Y \sim \text{Bin}(n,p)$. Here the mean value, $\text{E}(Y) = np$ and the variance is given by $\text{Var}(Y) = np(1-p)$. So the mean is related to the variance. The expected variance is known and does not need to be estimated separately as in the case of linear regression. Consequently, the z-score is used to test for the significance of each parameter instead of a t-score.

3.4. Confusion Matrix

Also known as an error matrix, it is used to visualize the performance of a classification algorithm. It is essentially a contingency table with two dimensions, actual and predicted. For instance, the columns of the matrix may represent the instances in each predicted class, while its rows may represent the instances in each actual class. The diagonal elements of the 2x2 matrix are the true positives and true negatives. Summing these two and dividing by the sum of all four elements of the matrix, yields the discriminatory power of a model (Everitt and Skrondal 2010, p. 99).

3.5. Akaike Information Criterion (AIC)

The AIC is an index used to to help choose among competing models. A smaller value of the index indicates the preferred model. The index is defined as follows:

$$\text{AIC} = 2k - 2\text{log}(\hat{L})$$

where k is the number of parameters and $\hat{L}$ is the maximum value of the likelihood function. The first term in the equation above penalizes a larger number of parameters; the more the parameters, the higher the AIC value. The second one takes into account the statistical goodness of fit of the model; the better the fit, the lower the index value (Everitt and Skrondal 2010, p. 10).

3.6. Backward Elimination Procedure

In regression analysis, it is necessary to choose a model that is neither underfittted nor overfitted. Backward elimination is a method used to select the most relevant covariates for predicting the response variable. We start by calculating the AIC for a model that includes all covariates. Next, each of the covariates $x_1, .., x_k$ is considered separately for removal from the model. First, x_1 is removed, and a new AIC value for the resulting model is calculated. Then x_1 is reintroduced, x_2 removed, and another AIC value is evaluated. This is repeated until x_k has been considered for removal. The AIC values of the models are then compared. If the AIC value of the model with all covariates is the least, the procedure stops. Otherwise, that covariate whose removal resulted in the least value of the AIC, is dropped from the model. This completes one iteration of the procedure. If a covariate has been dropped, the resulting model without that covariate is passed onto the next iteration of the procedure and the whole process is repeated. This continues until no further improvement in the AIC is achieved by dropping any single covariate (Everitt and Skrondal 2010, p. 386).

4. Statistical Analysis

4.1. Data Preprocessing

The response variable for the linear regression model is the rent per square meter of the properties. As this is not one of the 16 variables in the provided dataset, it has to be calculated by dividing each observation of the variable ‘totalRent’ by its corresponding observation of ‘livingSpace’ to form a new variable ‘sqmPrice’.

The variable ‘noParkSpaces’ assumes discrete values from 0 to 80. 7931 out of 12118 rental offers, or just above 65% of all observations have missing values for the number of associated parking spaces. For ease in interpretation, this variable is grouped into two categories. The category ‘0 or no information’ denotes missing values or the lack of a parking space and ‘1+’ denotes one or more parking spaces. This new transformed variable is named ‘parking_transf’.

The variable ‘typeOfFlat’ assumes 10 values. The value ‘apartment’ is made into its own category while the remaining 9 values are grouped into three more categories as follows: loft, maisonette, penthouse, terraced flat and ‘other’ are grouped to form the category ‘luxurious_artistic_other’; ground floor and raised ground floor are grouped to form the category ‘r_ground_floor’; and roof storey and half basement are grouped to form the category ‘roof_halfBasement’. Furthermore, for 714 observations there is no information about the type of flat. To avoid data loss, they are made into a separate category labelled ‘no information’. This transformation of the variable ‘typeOfFlat’ results in a new variable ‘type_flat_transf’ with five categories.

The variable ‘yearConstructed’ is transformed into a new variable ‘yearConstructed_transf’ by subtracting each observation from 2020. The transformed variable can be nicely interpreted as the age of a property in years. If ‘yearConstructed_transf’ is 0 or 1, the property was built in 2020 or 2019 and is considered to be newly constructed. For higher values of this variable, the property is 2 or more years old and is not regarded as being newly constructed.

The variable ‘regio2’ states the city in which a property is located. This has 54 distinct values in the dataset. This is transformed into a new variable ‘CityType’ by grouping the cities into three categories. The category ‘1-5’ indicates the 5 biggest cities of North Rhine-Westphalia by population size. These include Cologne, Dusseldorf, Dortmund, Essen and Duisburg. Similarly, the category ‘6-10’ indicates the next 5 biggest cities of the province which include Bochum, Wuppertal, Bielefeld, Bonn and Munster. The remaining cities are classified into a third category labelled ‘elsewhere’.

Henceforth, the variables ‘noParkSpaces’, ‘typeOfFlat’, ‘yearConstructed’ and ‘regio2’ are not utilized. Only their transformations are analysed and used to train both models. Moreover, the variable ‘ID’ has arbitrary values of the natural numbers and is not useful for the analysis. Therefore it is also dropped.

4.2. Univariate Analysis of the Variables

After the above transformations, the dataset now comprises of 6 categorical variables with 2 levels. One of these is the variable ‘parking_transf’ whose analysis shows that about 33% of the properties have one or more parking spaces. For around 67% of the rental offers, there is either no parking space or no information in this regard.

The other 5 nominal variables assume the values true or false. There are no missing values for any of these and they are summarized in table 1. Only 799 properties, or 6.6% of the rental offers are newly constructed. Hence, it is evident that there is an imbalance in the number of training examples for each category of the response variable in the logistic regression model. Besides this, table 1 also shows that around two-thirds of the properties have a balcony. Moreover, around 80% of them lack a fitted kitchen. About the same proportion of rental offers lack a lift and around the same percentage do not possess a garden.

Variable name	True	False
newlyConst	799	11319
balcony	8023	4095
hasKitchen	2438	9680
lift	2707	9411
garden	2140	9978

Table 1: Frequency distributions for five binary nominal variables

Average	Good	No information
5564	3709	2845

Table 2: Frequency distribution of the condition of the rental offers

Last 5 years	Over 5 years	No information
2872	1158	8088

Table 3: Frequency distribution of the year of last renovation

A+/A/B/C	D/E/F/G/H	No information
1419	2242	8457

Table 4: Frequency distribution of the energy efficiency class of the properties

Apartment	Luxurious/ artistic/other	Raised ground floor/ ground floor	Roof storey/ half basement	No information
7414	963	1315	1712	714

Table 5: Frequency distribution of the type of flat

1-5	6-10	Elsewhere
3280	1446	7392

Table 6: Frequency distribution of the city type in which the properties are located

There are 5 categorical variables that take on more than 2 levels. Table 2 shows how one of these, namely the condition of the properties, is distributed. Most of the properties (around 46%) are in an average condition. Likewise, tables 3 and 4, show the distributions for the year of last renovation and the energy efficiency class of the properties, respectively. As can be seen, for more than 8000 rental offers, there is no information regarding when they were last refurbished or in which energy efficiency class they lie. Lastly, tables 5 and 6 show how the flat type of the properties and the city type in which they are located is distributed. More than 60% of the properties are apartments and a similar number is located outside the 10 largest cities of North Rhine-Westphalia.

Figure 1 shows the frequency distribution of the floor in which the properties are located. As can be seen, most of them are situated in the first or second floor of a building. The first floor alone accounts for around 32% of the rental offers. Very few properties are located below ground floor or above the seventh floor.

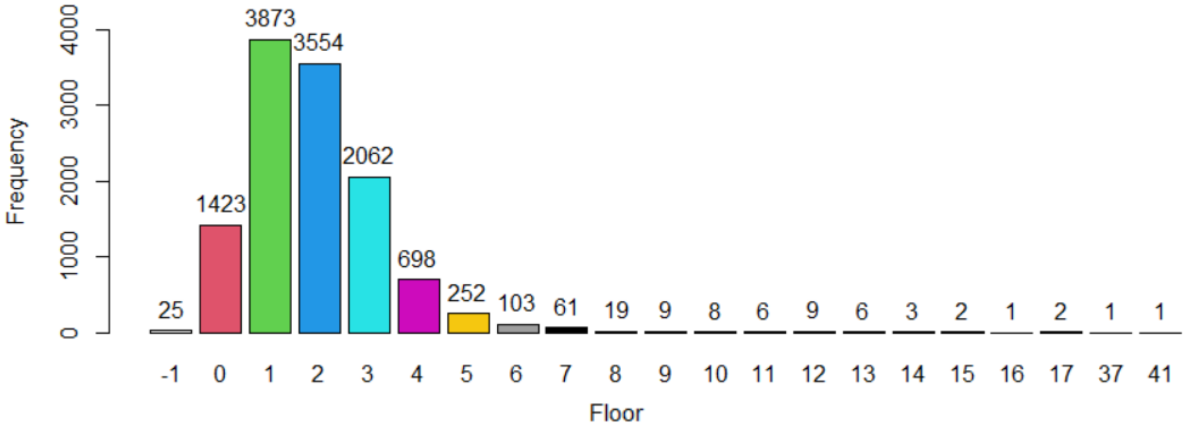

Figure 1: Frequency distribution of the floor in which the rental offers are located

The box plots for the total rent, rent per square meter, living space and the variable 'yearConstructed_transf' are shown in the appendix in figures 2, 3, 4 and 5 respectively. All four distributions are right skewed which is indicated by the points lying above the top whisker of each box plot. The median of the total rent is €670 and that of the rent per square meter is 10 €/m^2. The median size and age of the properties are 69 m^2 and 52 years respectively.

4.3. Results of the Linear Regression Model

For the linear regression model, the rent per square meter is modelled in terms of the remaining variables except the total rent. The 2284 observations with missing values for the total rent (and therefore also for the rent per square meter) are discarded from the training data. During dummy encoding the nominal variable 'condition', the category 'average' is taken as the reference. Similarly, for the variables 'lastRefurbish', 'energyEfficiencyClass', 'type_flat_transf', 'parking_transf' and 'CityType', the categories 'Last5Years', 'A+/A/B/C', 'apartment', '0 or no information' and '1-5' respectively, are taken as the references.

After setting up the model with all the covariates as described above, the backward stepwise procedure is carried out with the AIC as the model selection criterion. In the first iteration, the model with all covariates has an AIC value of 19620. Dropping the variables ‘garden’ or ‘energyEfficiencyClass’ from the model reduces this value to 19618 or 19619 respectively. Dropping any other covariate increases the index value. Therefore, after the first iteration the covariate ‘garden’ is dropped from the model.

In the second iteration, only the covariate ‘energyEfficiencyClass’ leads to a further reduction in the AIC value. Upon its removal, the index reduces from 19618 to 19617. Consequently, after the second iteration, this variable is also discarded from the model.

In the third iteration of the procedure, dropping any of the covariates only results in an increase in the value of the AIC. Therefore, no further covariate is dropped. The resulting model with the estimated parameter values is shown in table 7.

Covariate	Estimated parameter value	$(\Pr(>\|t\|)$
(Intercept)	12.835335	< 2e-16
newlyConstTRUE	1.639039	< 2e-16
balconyTRUE	0.157450	0.017522
hasKitchenTRUE	2.260416	< 2e-16
livingSpace	-0.019659	< 2e-16
liftTRUE	1.284051	< 2e-16
floor	0.045100	0.029395
conditiongood	1.068027	< 2e-16
conditionNO_INFORMATION	0.263851	0.000452
lastRefurbishNO_INFORMATION	-0.276472	7.89e-05
lastRefurbishOver5Years	-0.361290	0.000380
parking_transf1+	0.480581	1.25e-11
type_flat_transfluxurious_artistic_other	0.815811	4.84e-14
type_flat_transfnoinformation	0.014557	0.900438
type_flat_transfr_ground_floor	0.199272	0.041043
type_flat_transfroof_halfBasement	-0.454466	1.16e-07
yearConstructed_transf	-0.011117	< 2e-16
CityType6-10	-1.239836	< 2e-16
CityTypeelsewhere	-2.280656	< 2e-16

Table 7: Linear regression parameter estimates and their p-values

For $\alpha = 0.05$, the coefficients of the linear model are all statistically significant when tested for the hypothesis $H_0 : \beta_j = 0$ against $H_1 : \beta_j \neq 0$. The only exception is the dummy variable ‘type_flat_transfnoinformation’, which is 1 when there is no information about the type of flat of a property. Hence, there is insufficient evidence to suggest that changes in this particular variable are linked to changes in the response variable.

While interpreting any estimated coefficient, the values for the other ones are assumed to be constant. According to the trained linear model, for a property that has a balcony, a fitted kitchen or a lift, the rent per square meter is typically 0.16 €/m^2, 2.26 €/m^2 or 1.28 €/m^2 higher, respectively, than for a property lacking each of these features. So from the point of view of a renter for example, having a kitchen is more important than having a lift or a balcony to charge a higher square meter price.

Compared to a property in an average condition, the square meter rent for a property in a good condition is, in general, around 1.07 €/m^2 higher. If a property has been refurbished more than five years ago, then compared to a property refurbished in the past five years, its average square meter rent is around 0.36 €/m^2 lower. So if a property is in an average condition, by refurbishing it now to restore its good condition, we would expect its rent per square meter to increase by 1.43 €/m^2. However, the square meter price for a property constructed in 2019 or 2020 is, on average, 1.64 €/m^2 higher than for a property constructed earlier. Therefore, newly constructed properties are still more costly to rent than older ones that are recently renovated and in a good condition.

Unsurprisingly, as the figures suggest, rental properties in the 5 biggest cities of North Rhine-Westphalia are more expensive than those located outside of them. Moreover, out of all flat types, luxurious/artistic/other type of flats have the highest average square meter rent if other features are kept constant.

For every 10 m^2 increase in the size of a property, the rent per square meter typically decreases by 0.19 €/m^2. If the floor on which the property is located increases by 1, the square meter rent features an average increase of a meagre 0.05 €/m^2.

The adjusted R-squared value for the linear model is about 0.37. This means that it is able to explain only around 37% of the variation in the rental price per square meter of the properties. Figure 6 in the Appendix shows a scatter plot of the residuals of the model. The variation in the residuals seems to be much smaller for extreme fitted values compared to the fitted values in the center of the plot. Therefore, the assumption of constant variance appears to have been violated. Similarly, figure 7 shows the quantiles of the residuals plotted against theoretical normal quantiles. There is visible divergence from the straight line for higher quantiles indicating that the assumption of normally distributed errors may be compromised.

4.4. Results of the Logistic Regression Model

For the logistic regression model, the response variable ‘newlyConst’ is modelled in terms of the remaining variables except ‘yearConstructed_transf’ and ‘lastRefurbish’. Obviously, by directly knowing how old a property is or whether it was renovated in the past 5 years or not, predicting whether or not it was constructed in 2019/2020 becomes a trivial exercise. The 2284 observations with missing values for the total rent are substituted with the mean of the remaining observations. The rent per square meter is not used as a covariate as it is a derived variable that is linearly dependent on the total rent. The dummy coding of the variables is done in the same way as for the linear model in the previous subsection.

Once again, after setting up the model with all covariates as described above, the backward selection procedure is applied. The AIC value for the initial model is 3111. In the very first iteration, no variable is found whose removal would reduce this index value any further. Therefore, the procedure is terminated and no covariate is dropped from the model. The estimated values of the parameters are shown in table 8.

For $\alpha = 0.05$, three coefficients of the logistic model are statistically insignificant when tested for the hypothesis $H_0 : \beta_j = 0$ against $H_1 : \beta_j \neq 0$. These are associated with the dummy variables ‘balconyTRUE’, ‘type_flat_transfno information’ and ‘type_flat_transfroof_halfBasement’. For these dummy variables, there is insufficient evidence to conclude that a non-zero correlation exists between them and the response variable. In other words, there is not enough information to suggest that changes in these independent variables are actually associated with changes in the target variable ‘newlyConst’.

As in the last subsection, each coefficient is separately interpreted, assuming that the others are kept constant. The logistic model indicates that on average, the odds of a property being newly constructed, compared to the opposite case, increase by a factor of $e^{0.25} \approx 1.28$ if it has a balcony and by a factor of $e^{1.61} \approx 5$ if it has a lift. The same odds usually decrease to $e^{-1.37} \approx 0.25$ times their original value (or decrease by 4 times) if the property has a kitchen, and to $e^{-0.43} \approx 0.65$ times the original if it has a garden. So of these four binary nominal variables, knowing about the presence of a lift, is the most important factor for predicting if it has been constructed in the past two years.

Compared to a property in an average condition, a property in a good condition is, in general, a whopping $e^{4.35} \approx 77$ times more likely to be newly constructed than not. No

Covariate	Estimated parameter value	(Pr(>$\|z\|$))
(Intercept)	-6.7950830	$<$ 2e-16
balconyTRUE	0.2480002	0.095196
totalRent	0.0013079	$<$ 2e-16
hasKitchenTRUE	-1.3733649	$<$ 2e-16
livingSpace	-0.0145392	6.39e-09
liftTRUE	1.6119625	$<$ 2e-16
floor	-0.1902550	1.78e-06
gardenTRUE	-0.4267352	0.000874
conditiongood	4.3502561	$<$ 2e-16
conditionNO_INFORMATION	2.2841652	1.04e-08
energyEfficiencyClassD/E/F/G/H	-3.0874965	2.67e-11
eenergyEfficiencyClassNO_INFORMATION	-0.6545876	5.03e-10
parking_transf1+	0.9426998	$<$ 2e-16
type_flat_transfluxurious_artistic_other	0.4641897	0.000818
type_flat_transfnoinformation	0.1331937	0.541442
type_flat_transfr_ground_floor	0.3077474	0.037367
type_flat_transfroof_halfBasement	0.0022905	0.990226
CityType6-10	0.4126344	0.015591
CityTypeelsewhere	0.6755490	2.06e-07

Table 8: Logistic regression parameter estimates and their p-values

other independent variable in the model features such disproportionate odds. So this covariate is of paramount importance when it comes to aiding in anticipating whether a flat is newly constructed or not.

The average odds of being newly constructed for a property with a D/E/F/G/H energy efficiency class are $e^{-3.09} \approx 0.05$ times the odds of being newly constructed for a property whose rating is A+/A/B/C. In other words, if a property has an energy efficiency class of A+/A/B/C compared to D/E/F/G/H, then its odds of being newly constructed increase by a factor of $1/0.05 = 22$.

Since the dummy variable associated with the category luxurious/artistic/other has the largest coefficient of all variables linked to the type of flat, these kinds of buildings are more likely to be constructed in the last two years than any other class. For example, compared to an apartment, the average odds of a property being newly constructed increase, by a factor of $e^{0.46} \approx 1.58$ if it is of the type luxurious/artistic/other.

Furthermore, compared to a property located in the five biggest cities of North Rhine-Westphalia, if a property is located in the 6 to 10 largest cities of province, the average odds of being newly constructed increase by a factor of $e^{0.41} \approx 1.51$; and if located elsewhere, they increase by a factor of $e^{0.67} \approx 1.95$. This means that newly constructed properties are more likely to be found in smaller cities of the province. Finally, a 10 m^2 increase in the living space of a property usually decreases the odds of being newly constructed odds to $e^{0.015 \cdot 10} \approx 0.86$ times their original value.

		Predicted	
		Yes	No
Actual	Yes	348	451
	No	216	11103

Table 9: Confusion Matrix for the Logistic Model

The confusion matrix for the logistic regression model is shown in table 9. The sum of the table's diagonal entries, which are the true positives and true negatives, is 11451. This is around 94.5% of the total observations. So the trained logistic model has a very high overall discriminatory power. However, there is a large difference in the number of training examples of the two outcomes. Only 799 or approximately 6.6% of all training examples comprise of newly constructed rental properties. The top row of table 9 shows that out of these, only 348 or about 43.5% are correctly classified. This shows that while the trained model has a low overall classification error, it performs poorly on training examples where the properties are newly constructed.

5. Summary

This report concerns the rental price data for the province of North Rhine-Westphalia, Germany. 12118 rental offers are extracted from the real estate web-portal Immobilienscout24. For each offer, the following 16 variables are provided in the dataset: its ID; its total rent in Euros; its year of construction; the year when it was last renovated; the number of parking spaces provided with it; its size in square meters; whether or not it is newly constructed (i.e. constructed in the year 2019 or 2020); whether or not it has a balcony, a kitchen, a lift or a garden; its type of flat; the floor in which it is situated; the city in which it is located; its condition; and its energy efficiency class.

The dataset is preprocessed to deal with missing values in some of the variables, to group some of them into categories and to transform some of them for ease of interpretation. Thereafter, a univariate analysis of the variables is done. This reveals that most houses have a balcony but lack a kitchen, a lift or a garden. The majority of the rental offers are in an average condition. There are a large number of missing values for the year of last renovation and energy efficiency class. Most of the buildings are apartments; most are situated in the first or second floor of buildings; and most are located outside the ten largest cities of North Rhine-Westphalia. The properties are around 50 years old on average. They have a median rent of €670 and a median size of around 70 m^2.

After preprocessing and univariate analysis, a linear regression model is trained with the rent per square meter as the response variable. This is calculated using the total rent and the size of each property. The remaining variables except total rent are used as covariates. Using a backward elimination procedure and the AIC as the model selection criterion, the variables garden and energy efficiency class are discarded from the model. The resulting model is able to explain just 37% of the variation in the response variable. The assumptions of the linear regression model do not seem to be met.

Afterwards, a logistic regression model is built to predict whether a property is newly constructed or not. The remaining variables are used as covariates, except for the year of construction and the year of last renovation. The backward elimination procedure using the AIC does not result in the removal of any covariate. The confusion matrix for the trained model shows that it has an overall discriminatory power of above 94%. However, it only classifies less than 45% of the newly constructed rental offers correctly. This is perhaps explained by the fact that very few, i.e. less than 7% of all training examples are newly constructed properties.

To improve the quality of the results, the size of the dataset should be increased. Especially for the logistic regression model, more newly constructed properties should be used during training. The variables could be classified into different categories to yield different models. Polynomial effects as well as interactions between the variables can also be incorporated into the model equations. To get a better estimate of the generalization error, the data should be divided into training and test data and e.g. a 10-fold cross-validation could be used. The experiment can be extended by incorporating more variables into the models like the crime rate of the area where a property is located, etc.

References

[1] B. S. Everitt and A. Skrondal. *The Cambridge Dictionary of Statistics, Fourth Edition*. Cambridge University Press, 2010.

[2] Ludwig Fahrmeir et al. *Regression: Models, Methods and Applications*. Jan. 2013. ISBN: 978-3-642-34332-2. DOI: `10.1007/978-3-642-34333-9`.

[3] Pladson Kristie. *Stuttgart unseats Munich as Germany's most expensive city for renters*. Deutsche Welle. Nov. 2019. URL: `https://www.dw.com/en/stuttgart-unseats-munich-as-germanys-most-expensive-city-for-renters/a-51374468` (visited on 01/30/2021).

[4] R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria, 2020.

[5] Amt für Wohnen und Migration München. *Mietspiegel für München*. 2019. URL: `https://www.muenchen.de/rathaus/Stadtverwaltung/Sozialreferat/Wohnungsamt/Mietspiegel.html` (visited on 01/30/2021).

A. Additional figures

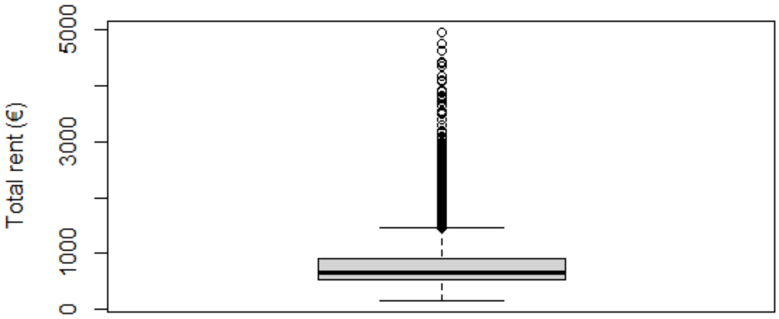

Figure 2: Box plot of the total rent of the properties

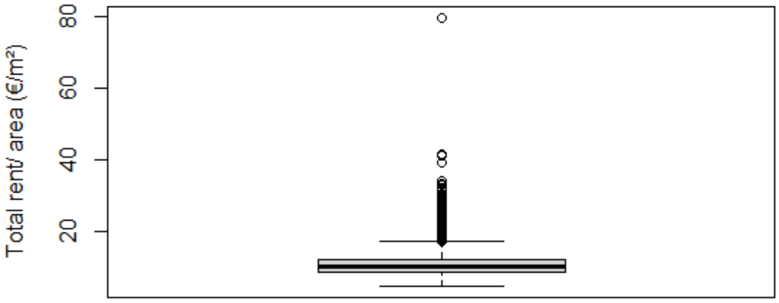

Figure 3: Box plot of the total rent per square meter of the properties

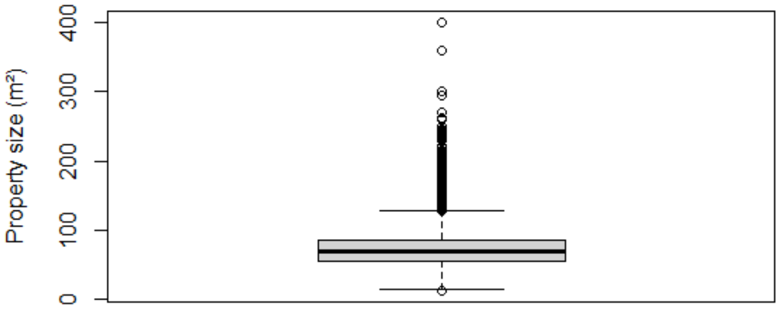

Figure 4: Box plot of the size of the properties

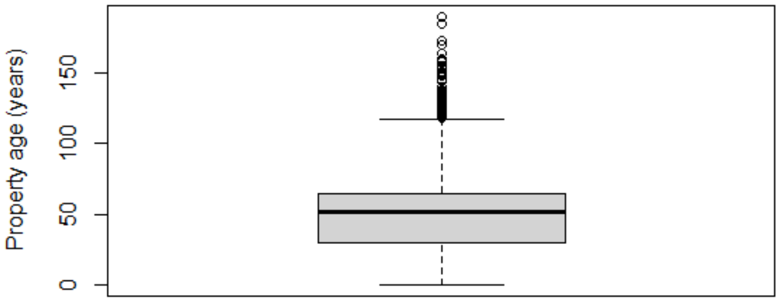

Figure 5: Box plot of the variable ‘yearConstructed_transf’

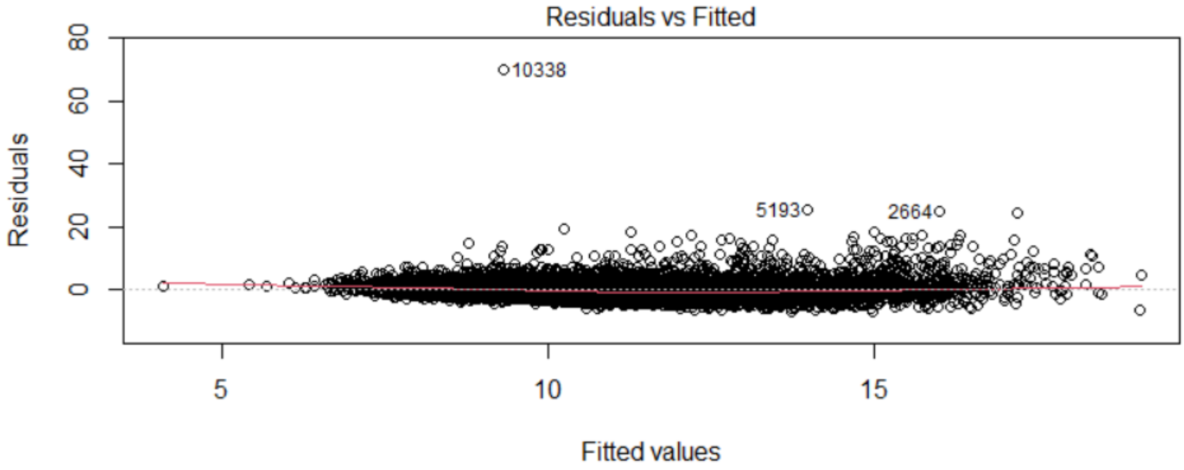

Figure 6: Scatter plot of the residuals of the fitted linear model

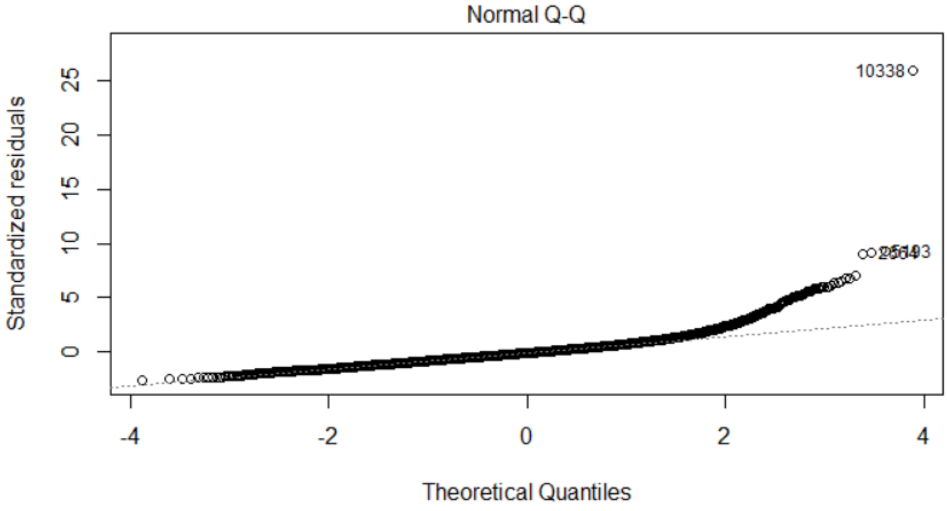

Figure 7: Q-Q plot of the residuals of the fitted linear model against normal theoretical quantiles