TU Dortmund

Introductory Case Studies

Project 3: Regression Analysis

Lecturers:
Prof. Dr. Jörg Rahnenführer
Dr. Philipp Adämmer
Dr. Andrea Bommert
M. Sc. Hendrik Dohme

Author: Alp Yalçın

Group number: 3

Group members: Ezgi Osmanoğlu, Elif Aykut, Yazan Alhalabi, Marges Pinderi

July 8, 2021

Contents

1 Introduction

Rental properties and their prices is one of the riveting topics in society. Rental property market has never lost in value since years and lots of parameters affect on the prices of rental properties.

The main purpose of this report is modelling the price of rents in Dortmund city. In this project, 16 different parameters are used to estimate the price of rent per square meter.

In the second section, dataset and data quality is explained. Observation number and structure of the parameters are clarified. Later, in third section, statistical methods which are used in statistical analysis part are expressed in detailed. Linear regression, Akaike information criteria, Bayesian information criteria and confidence intervals explained.

In the fourth section, data preparation is made for starting the regression analysis. N/A values are excluded and observations are filtered for working on Dortmund city observations. Flat type variable is grouped into 4 categories. Dependent variable is chosen as rental price per square meter and computed. Then, Best subset of predictors are found and different selection criteria are used and results are compared. Best linear model is estimated and confidence intervals are provided for the regression parameters. Additionally, goodness of fit is evaluated.

Finally, the fifth section summarizes the main results and some comments are given about the project.

2 Problem statement

2.1 Data set and data quality

This report is compiled from (Bartelheimer, 2020) (immoscout24.de) web-portal which is one of the biggest real estate in Germany. Immobilienscout24 has records for both rental properties and homes for sale, nevertheless, dataset only includes rental properties.

The dataset includes 12 118 observations of rental offers for properties which are located in the province North Rhine- Westphalia as of February 20, 2020. The data set also contains 16 variables for these rental offers. It includes 4 integer variables which are identification number (ID), construction year (yearConstructed), number of parking spaces provided with property (noParkSpaces) and the floor the property is in (floor). Additionally, it has 5 logical (Boolean) variables which are whether the property is newly constructed, (newlyConst), has a balcony (balcony), has a kitchen (hasKitchen), has a lift (lift), and has a garden (garden) or not. Also, it contains 5 character variables which are type of the flat (typeOfFlat), city/municipality where the property is located (regio2), condition of the property (condition), year of last renovation (lastRefurbished) and energy efficiency class of the building (EnergyEfficiencyClass). Last but not least, dataset contains 2 numeric variables which are total rent (totalRent) and property size in square meters (livingSpace).

Only Dortmund city region observations are used in the report and the size of the dataset for Dortmund includes 468 observations.

Overall, data quality is adequate and data set is relatively clean. However, 3 variables contain N/A values which are noParkSpaces variable with 420 N/A values, totalRent variable with 75 N/A values and typeOfFlat variable with 21 N/A values. All N/A values of the dataset and noParkSpaces variable which contains lots of N/A values are excluded from the analysis of this report.

2.2 Project objectives

The main objective of this report is making regression analysis of the dataset with the given variables and observations. Therefore, as a first objective, data preparation is applied. This project is only interested in Dortmund city rental properties observations, so first of all, observations for the city of Dortmund is selected. Additionally, all of the rows which has N/A values are omitted and variable which has the highest number of missing observations is removed. Secondly, rental price per square meter (sqmPrice) is computed and added as a dependent variable for the regression analysis. Thirdly, typeOfFlat variable is separated to 4 categories. As a second objective, linear regression is used. Firstly, linear model is used with all variables. All regressions are fitted and all possible subsets of the set of potential independent variables are tested. Best predictors for sqmPrice is found with using Best Subset Selection and Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC) are used as selection criteria. Also included variables are compared for two models. Secondly, best linear model for sqmPrice using the AIC is estimated. Coefficients of the model and their statistical significance are interpreted and confidence intervals for the regression parameters are provided and evaluated the goodness of fit.

3 Statistical methods

For all analyses and calculation, the software (R Development Core Team, 2020) with packages **dplyr** (Wickham et al., 2021), **olsrr** (Hebbali, 2020), **gmodels** (Warnes et al., 2018), **ggplot2** (Wickham, 2016), **xtable** (Dahl et al., 2019), **broom** (Robinson et al., 2021)

3.1 Linear Regression

Linear regression is a basic and commonly used type of predictive analysis. The idea of regression is to examine how predictive variables perform in predicting a dependent variable and which variables are significant predictors of the outcome variable. These regression estimates are used to describe the relationship between one dependent variable and one or more independent variables.

(Fahrmeir et al., 2007, p. 74)

$y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k + \varepsilon$

(Fahrmeir et al., 2007, p. 74)

Where $x_1, x_2...$ represents the covariates, k the number of covariates, $\beta_0, \beta_1, \beta_2, ...\beta_k$ typifies the coefficients and ε represents the error term. The covariates can be continuous, binary, or multi-categorical.

(Fahrmeir et al., 2007, p. 74)

The model $\mathbf{y}$=$\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ is called the classical linear model representation, where n represented as sample size in that matrix notation below;

$$\mathbf{y} = \begin{pmatrix} y_1 \\ . \\ . \\ y_n \end{pmatrix}, \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ . \\ . \\ \beta_k \end{pmatrix}, \mathbf{X} = \begin{pmatrix} 1 & X_{11} & ... & X_{1k} \\ . & . & . & . \\ . & . & . & . \\ 1 & X_{n1} & ... & X_{nk} \end{pmatrix}, \boldsymbol{\epsilon} = \begin{pmatrix} \epsilon_1 \\ . \\ . \\ \epsilon_n \end{pmatrix}$$

(Fahrmeir et al., 2007, p. 75)

Classical Linear Model Assumptions

1.The errors have mean or expectation zero $E(\boldsymbol{\epsilon}) = 0$

2.It is assumed that fixed error variance σ^2 along observations, which is homoscedastic with $Var(\epsilon_i) = \sigma^2$. When the variances diverge along observations, the errors are named heteroscedastic, $Var(\epsilon_i) = \sigma_i^2$ We postulate that errors are uncorrelated $Cov(\epsilon_i, \epsilon_j) = 0$ for $i \neq j$ Therefore, the covariance matrix $Cov(\boldsymbol{\epsilon}) = E(\boldsymbol{\epsilon\epsilon'}) = \sigma^2\boldsymbol{I}$

3.We assume that design matrix $\boldsymbol{X}$ has full column rank $rk(\boldsymbol{X}) = k + 1 = p$

4. We presume a normal distribution for the errors to construct confidence intervals and hypothesis tests for the regression coefficients. With first two assumptions, we attain $\boldsymbol{\epsilon} \backsim N(\boldsymbol{0}, \sigma^2\boldsymbol{I})$

(Fahrmeir et al., 2007, p. 75)

Covariate only effect the mean of y. The variance σ^2 of y_i or the covariance matrix $\sigma^2\boldsymbol{I}$) of y is independent of covariates.

$\boldsymbol{y} \backsim \boldsymbol{N(X\beta}, \sigma^2\boldsymbol{I})$

(Fahrmeir et al., 2007, p. 75)

Parameter Estimation in the Classical Linear Model

The theory of estimation for the regression coefficients of linear models is closely bounded to the least squares method, founded by Legendre in 1806.Additionally, other principles for parameter estimation are possible.

(Fahrmeir et al., 2007, p. 104)

According to the principle of least squares, the unknown regression coefficients are estimated by minimizing the sum of the squared deviations, it contains in minimizing the sum of square deviations with relation to the β coefficients. (Fahrmeir et al., 2007, p. 104)

In the classical linear model, the estimator

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

where $\boldsymbol{\beta}$ is estimated coefficients, $\mathbf{X}$ represents the observation and $\mathbf{y}$ signify as dependent variable, minimizes the least squares criterion

(Fahrmeir et al., 2007, p. 105)

$$LS(\boldsymbol{\beta}) = \sum_{i=1}^{n} (y_i - x_i^{'}\beta)^2 = \sum_{i=1}^{n} \epsilon_i^2 = \boldsymbol{\epsilon}^{'}\boldsymbol{\epsilon}$$

Under the assumption of normally distributed errors, the least squares estimator is also the ML estimator for $\boldsymbol{\beta}$.

Estimation of the variance error is :

$$\hat{\sigma}^2 = \frac{\hat{\boldsymbol{\epsilon}}\boldsymbol{\epsilon}}{n - p^{'}}$$

Where $\hat{\epsilon}$ is the residual of the model.

(Fahrmeir et al., 2007, p. 105)

Predicted Values and Residuals

Based on the least squares estimator, the conditional mean of $\mathbf{y}$ can be estimated by $\hat{\boldsymbol{y}} = \mathbf{X}\hat{\boldsymbol{\beta}}$ and with substituting the least squares estimator further results in
$\hat{\boldsymbol{y}} = \mathbf{X}(\mathbf{X}^{'}\mathbf{X})^{-\mathbf{1}}\mathbf{X}^{'}\mathbf{y} = \mathbf{Hy}$ with the nxn-matrix $\mathbf{H} = \mathbf{X}(\mathbf{X}^{'}\mathbf{X})^{-\mathbf{1}}\mathbf{X}^{'}$

(Fahrmeir et al., 2007, p. 107)

The matrix $\mathbf{H}$ which is symmetric and idempotent , is also called the prediction matrix or hat matrix. It is also possible to express the residuals in matrix notion with

$\hat{\boldsymbol{\epsilon}} = \mathbf{y}$ - $\hat{\boldsymbol{y}} = \mathbf{y}$ - $\mathbf{Hy} = (\boldsymbol{I}$- $\mathbf{H})\mathbf{y}$

(Fahrmeir et al., 2007, p. 107)

Standardized Residuals

Although residuals frequently used to affirm model assumptions in linear model, the residuals are not always sufficient for this purpose. The residuals are neither homoscedastic nor uncorrelated. In heteroscedasticity of residuals,correlation can not be neglectable.

Therefore, the justification of the assumption of homoscedastic errors is problematic, because heteroscedastic residuals are in general not in indicator for heteroscedastic errors. Standardization is the efficient solution of the heteroscedasticity issue. We attain the standardized residuals with dividing through the estimated standard deviation of residuals.

(Fahrmeir et al., 2007, p. 124)

$$r_i = \frac{\hat{\epsilon_i}}{\hat{\sigma}\sqrt{1 - h_{ii}}}$$

Standardized residuals are homoscedastic with the proivded that the model assumptions are correct. Therefore, analysis of standardized residuals assists to examine whether or not the assumption of homoscedastic variances is outraged

(Fahrmeir et al., 2007, p. 124)

h_{ii} is the ith diagonal element of the hat matrix.

Estimated coefficients significance tested with t-test with the significance level of 0.05.

Test of significance (t-test): $H_0 : \hat{\beta}_j = 0$ against $H_1 : \hat{\beta}_j \neq 0$

Test can be based on the "t-statistics,"

$t_j = \frac{\hat{\beta}_j}{se_j} \sim t_{n-p,(1-\alpha/2)}$

where $s\hat{e}_j = Var(\hat{\beta_j})^{1/2}$ denotes the estimated standard deviation of standard error of $\hat{\beta}_j$ which can be found as the diagonal elements of the matrix $cov(\hat{\beta}) = \hat{\sigma}^2(\mathbf{X'X})^{\mathbf{-1}}$ (Fahrmeir et al., 2007, p. 124)

Confidence Intervals

The duality between two-sided tests and confidence intervals regions construct a confidence interval for a parameter.

Probability of not rejecting H_0 is;

$$[\hat{\beta}_j - t_{(n-p)}(1-\alpha/2)s\hat{e}_j, \hat{\beta}_j + t_{(n-p)}(1-\alpha/2)s\hat{e}_j]$$

as a $(1-\alpha)$ confidence interval for β_j Confidence interval is %95 because signicance level is $\alpha = 0.05$
(Fahrmeir et al., 2007, p. 136)

Dummy Coding for Categorical Covariates

For modelling the effect of a covariate $x \epsilon$ 1,....c with c categories using dummy coding, we define the $c-1$ dummy variables

(Fahrmeir et al., 2007, p. 97)

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

for $i = 1, .., n$, and include them as explanatory variables in the regression model $y_i = \beta_0 + \beta_1 x_{i1} + ... + \beta_{i,c-1} x_{i,c-1} + ... + \epsilon_i$ where i represent the respective observation We exclude one of the dummy variables for identifiability, for this case the dummy variable for category c which is reference category. By direct comparison with reference category,estimated effects can be interpreted.

(Fahrmeir et al., 2007, p. 97)

Coefficient of Determination

The coefficient of determination is closely related to the empirical correlation coefficient and can be used as a goodness-of-fit measure. (Fahrmeir et al., 2007, p. 112)

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

The closer the coefficient of determination is to 1, the smaller is the residual sum of squares, and the better the fit is to the data. If $R^2 = 1$, all residuals are zero with perfect fit to the data.

Comparing the coefficient of determination across models is meaningful for common dependent variable and if the models which contain the same number of parameters and intercept.

(Fahrmeir et al., 2007, p. 114)

The Corrected Coefficient of Determination

Coefficient of determination is restricted when comparing different models because it will always increase and never decrease with adding new covariate into the model. Including a correction term for the number of parameters is the solution for this problem which is called corrected coefficient of determination. It is defined by

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

(Fahrmeir et al., 2007, p. 148)

Best Subset Selection Spearate least squares are fitted for each possible combination of p predictors to perform best subset selection. All p models are fitted which comprise exactly one predictor, all $p * (p-1)/2$ models which contain exactly wto predictors, and so on. Then after checking all results, best subset is identified.

(James et al., 2013, p. 205)

3.2 Akaike information criterion(AIC)

The Akaike information criterion (AIC) is one of the most widely used criteria for model choice within scope of likelihood-based inference. AIC is defined by

$$AIC = -2.l(\hat{\boldsymbol{\beta_M}}, \hat{\sigma}^2) + 2(|M| + 1)$$

Where the value of $l(\hat{\boldsymbol{\beta_M}}, \hat{\sigma}^2)$ is the maximum log likelihood when the estimators of the model parameters $\hat{\boldsymbol{\beta_M}}$ inserted, the error variance $\hat{\sigma}^2$. The smaller values of the AIC, means the model fitted better.

$M + 1$ value is the total number of parameter in model. In linear model with Gaussian Errors, the ML is given by:

$$-2.l(\hat{\boldsymbol{\beta_M}}, \hat{\sigma}^2) = n.log(\hat{\sigma^2})$$

Therefore,

$$AIC = n.log(\hat{\sigma^2}) + 2(|M| + 1)$$

where n is the observation number.

(Fahrmeir et al., 2007, p. 148)

3.3 Bayesian information criterion(BIC)

The form of the BIC is similar to that of the AIC, and again smaller values indicate a better model fit. Note, however, that the BIC and AIC are motivated in a very different way. From a practical point of view, the main difference is that the BIC penalizes complex models much more than the AIC. Thus, the resulting “best” models are typically more parsimonious when using the BIC rather than the AIC.

The Bayesian information criterion (BIC) is described with;

$BIC = -2.l(\hat{\boldsymbol{\beta}}_{\boldsymbol{M}}, \hat{\sigma}^2) + log(n)(|M| + 1)$

Assuming Gaussian errors, we obtain;

$BIC = n.log(\hat{\sigma}^2) + log(n)(|M| + 1)$

(Fahrmeir et al., 2007, p. 150)

4 Statistical analysis

In this section, the statistical methods explained above are applied to the presented data set and the results are interpreted.

4.1 Data Preparation

Data preparation part is one of the most significant parts of this project to make the regression analysis of this real estate rental dataset. Cleaning the data and understanding the structure of the dataset is crucial for this report.

4.1.1 Selection and Data Cleaning

Initially, structure of the dataset is analysed and 12118 observations and 16 variables are obtained in the dataset. Dataset has 4 integer variables, 5 logical(Boolean) variables, 2 numeric variables and 5 character variables. As a first task of the report, Dortmund city region is filtered from the regio2 variable because this report is only focused on Dortmund city regression analysis. As a result, the size of observations is decreased more than 20 times than the original size of the dataset after filtering only the Dortmund city region. The observation number is decreased to 558 observations from 12118. Secondly, all variables which have N/A values are examined. 3 variables which are noParkSpaces, totalRent and typeOfFlat has N/A values with the numbers of 420, 75 and 21 missing observations, respectively. Due to having highest number of N/A values among the variables, noParkSpaces variable is entirely excluded from the dataset. After excluding the noParkSpaces, 468 observations are contained left in the dataset. After excluding the noParkSpaces variable, all of the rows which have at least one N/A values are also excluded from the dataset. The rows which contain "NO INFORMATION" is still kept in the dataset. After omitting the rows which have N/A values, 467 observations and 15 variables is contained in the dataset.

4.1.2 Computation of rental price per square

Thirdly, rental price per square meter (sqmPrice) which is computed to be used as a dependent variable in the regression analysis, is calculated with dividing the livingSpace variable from totalRent continuous variable and is added to the dataset.

4.1.3 Grouping type of flats into four categories

As a final step of the data preparation part of the dataset, typeOfFlat variable which contains ten unique values is grouped into four categories.

These categories are "apartment" which comprises only apartment values, "luxurious artistic other" which comprises the values loft, maisonette, penthouse,terrace flat and other values, "r ground floor" that comprises ground floor and raised ground floor values, "roof half basement" that comprises the values roof storey and half basement values. The value numbers of these categories are 357, 25, 40 and 46, respectively.

4.2 Linear Regression

Report is focused on two different tasks on linear regression part of statistical analysis chapter. The reports main objective for this part is finding the best predictors for sqmPrice with using different selection criteria and compare them. Additionally, the best linear model is estimated with using the Akaike Information Criterion and that linear model is examined with respect to different perspectives.

As a beginning, 3 variables of the dataset are excluded from the dataset which are ID, totalRent and regio2. ID variable is excluded since it has no effect on the regression analysis. Additionally, totalRent is excluded because that variable is used to calculate our dependent variable rental price per square meter (sqmPrice) and it should be omitted for not to effect on results. The other variable, livingSpace, which is used to determine sqmPrice is still included because it has no defective effect on the regression result. Last but not least, Regio2 variable is also excluded because it only contains Dortmund city observations and does not needed in the dataset as a variable column.

Descriptive Analysis for Numeric Variables

	yearConstructed	livingSpace	sqmPrice
X	Min. :1898	Min. : 13.00	Min. : 6.308
X.1	1st Qu.:1955	1st Qu.: 54.00	1st Qu.: 9.230
X.2	Median :1966	Median : 67.78	Median :10.353
X.3	Mean :1967	Mean : 69.57	Mean :10.614
X.4	3rd Qu.:1976	3rd Qu.: 79.99	3rd Qu.:11.507
X.5	Max. :2020	Max. :187.00	Max. :18.975

The oldest building in the dataset is built in 1898. Also the median number of constructed year is 1966, where the mean is 1967. The length living space of the buildings ranges between 13 m^2 and 187 m^2. Also the mean is 69.57 m^2 and and the median is 67.78 m^2. Rental price per square ranges between 6.308 and 18.975 whereas the mean is 10.614 per square.

Descriptive Analysis for Categorical Variables

First of all, linear model is implemented with all 13 variables included which will be used in regression analysis of the report. Type of flat variable values are grouped into 4 different categories apartment, luxurious artistic order, r ground floor, roof half basement with the numbers of 357, 25, 40, 46 observations, respectively, on the data preparation part of the statistical analysis. Apartment category is chosen as a reference category for that variable in linear model.

Additionally, for values which named as ”NO INFORMATION”, we do not have precise information. However, as it is mentioned on the data preparation part again, we decide to include rows which has NO INFORMATION” values on the dataset. Therefore, we consider the “NO INFORMATION” values as a category in our analysis.

energyEfficiencyClass variable has 3 unique values which are "Aplus/A/B/C", "D/E/F/G/H" and "NO INFORMATION" with 50, 69, 349 observation numbers, respectively. Aplus/A/B/C" category is chosen as a reference category for that variable in linear model. Condition variable has 3 categories with 107 “good” observations, 174 “average” observations and 187 “NO INFORMATION” observations. For this variable, “average” category is chosen as reference category. Additionally, LastRefurbish variable has also 3 categories which are "NO INFORMATION" "Over5Years" and "Last5Years" with numbers of 325, 50 and 93 observations, respectively. “Last5Years” is chosen as a reference category for the linear model. Floor variable ranges between -1 floor to 11 floor, most of the observations are fall between 0 and 3 floors.

For binary variables, “newlyConst” has 451 false and 17 true observations, “balcony” has 139 false and 329 true, “hasKitchen” has 387 false and 81 true, “lift” has 372 false and 97 true observations and finally, “garden” has 419 false and 49 true observations.

4.2.1 Finding Best Predictors for Dependent Variable

First of all, linear model is used with all 12 features in the dataset. The result shows that including kitchen and lift have significant positive impact on price of square meter of rental properties in Dortmund. Additionally, the properties condition is and whether the property is newly constructed has positive effect on price per square meter. On the other hand, interestingly, the property size has significantly negative influence on price per square.

After liner model is implemented with all predictors in the dataset, all possible combination of the predictor variables is applied with the best subset regression model to select the best model according to Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC).

According to Akaike Information Criterion (AIC) score, the best linear model can be found with 9 predictor variables which are newlyConst, yearConstructed, hasKitchen,livingSpace, lift, typeOfFlat, floor, condition and energyEfficiencyClass with the score of 1908.058.

On the other hand, according to Bayesian Information Criterion (BIC) score, with the score of 1956.386, the best linear model can be reached with 6 independent vatiables which are newlyConst, yearConstructed, hasKitchen,livingSpace, lift and condition.

Criteria	Model	Value
AIC	newlyConst yearConstructed hasKitchen livingSpace lift typeOfFlat floor condition energyEfficiencyClass	1908.0583
BIC	newlyConst yearConstructed hasKitchen livingSpace lift condition	1956.3864

Bayesian Information Criterion (BIC) penalizes the parameters more strictly than Akaike Information Criterion (AIC). Therefore, in our dataset, for the best prediction selection, 3 predictors which are type of the property, the floor the property is in and energy efficiency class of the building, are included in AIC but are excluded in BIC.

4.2.2 Estimate the “best” linear model for sqmPrice using the AIC

At this section of the statistical analysis part, the "best" linear model for sqmPrice using the AIC is estimated with 9 parameters which are explained in previous part.

First of all, assumptions for classical linear regression model is checked. For the homoscedasticity assumption, standardized residual plot is checked. On the y axis, predictions (Eur/m^2) plotted and on the x axis Residuals (Eur/m^2) are shown. According to the plot we have, we can hold this assumption.

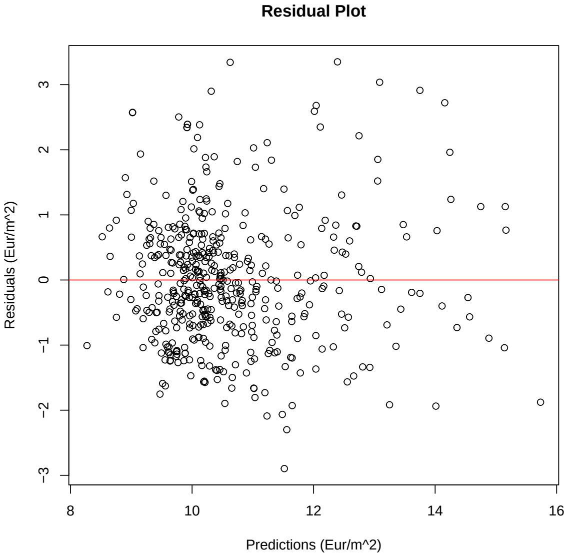

Figure 1: Residual Plot

For normality assumption, we used QQ-Plot with standardized residuals.On the y axis, standardized residuals plotted and and on the x axis theoretical quantiles is ranged. We can also say that this assumption holds true too.

Secondly, coefficients of the model and their statistical significance are examined and confidence intervals are provided for the regression parameters.

Again the results show that, having kitchen and lift, being in good condition, also the construction year and the being newly constructed features have significant effect on prices per m^2

Also it is obvious to see that, when the other parameters are fixed, for the 5 of the parameters which are type of flat, roof half basement, floor, condition with no information category and energy efficiency class D/E/F/G/H, we can not reject the null hypothesis.

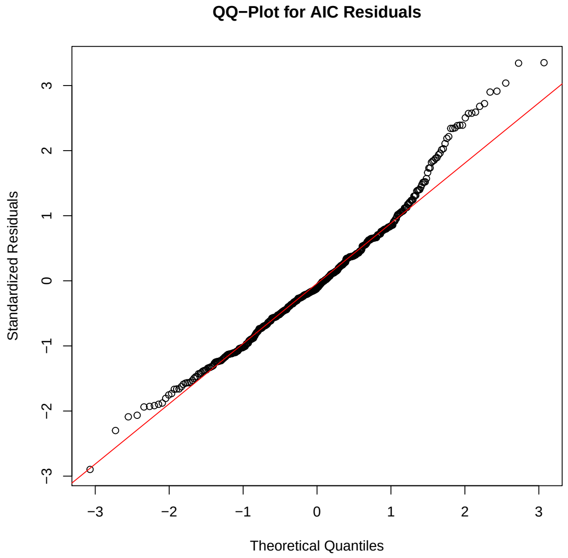

Figure 2: QQ Plot

Therefore we can interpret that these parameters have no impact on the square meter rental prices in m^2

	Estimate	CI lower	CI upper	Std. Error	p-value
(Intercept)	-5.38	-20.21	9.45	7.55	0.48
newlyConstTRUE	1.61	0.59	2.63	0.52	0.00
yearConstructed	0.01	0.00	0.02	0.00	0.02
hasKitchenTRUE	0.92	0.44	1.40	0.25	0.00
livingSpace	-0.02	-0.03	-0.01	0.00	0.00
liftTRUE	0.79	0.29	1.29	0.26	0.00
typeOfFlatluxurious_artistic_other	0.83	0.01	1.65	0.42	0.05
typeOfFlatr_ground_floor	-0.68	-1.33	-0.03	0.33	0.04
typeOfFlatroof_half_basement	-0.49	-1.10	0.11	0.31	0.11
floor	-0.13	-0.27	0.01	0.07	0.07
conditiongood	1.19	0.71	1.68	0.25	0.00
conditionNO_INFORMATION	-0.14	-0.56	0.28	0.21	0.51
energyEfficiencyClassD/E/F/G/H	-0.68	-1.37	0.01	0.35	0.05
energyEfficiencyClassNO_INFORMATION	-0.78	-1.34	-0.21	0.29	0.01

Finally, goodness of fit is evaluated with r squared and adjusted r squared measures are used. With the score of 0.30, we can say that % 30 of the variation in the output variable is explained by the input variables. Our model parameters are not encounter the dependent variable well.

	adj.r.squared	r.squared	AIC	BIC	p.value
1	0.30	0.32	1908.06	1970.29	0.00

5 Summary

In this project, (Bartelheimer, 2020) Immobilienscout24 real estate web-portals data set was used to modelling the price of the rent in Dortmund. The given dataset had 12 118 observations and 16 variables, the places which are located in North Rhine-Westphalia region. First of all , data was cleaned and N/A values are excluded. Living space and total rent variables were used to compute the dependent variable which was selected as price per square meter. After that calculation, to avoiding the correlation with dependent variable, total rent variable was excluded from the dataset too.

After preparation of the dataset, linear model was implemented with all predictors and their impacts on the model was interpreted. After, all possible combinations of the predictor variables are applied with the best subset regression model to select the best model, AIC and BIC scores was examined. Both criterions chose whether the properties constructed new or not, condition of the property, whether it has kitchen and lift, and also chose the energy efficiency class as a parameters. BIC criteria selected less parameters because of its strict penalization structure. Interestingly, living space showed that that parameter had negative impact on price per square which we did not expect before making the analysis. The result showed that people prefer smaller properties to rent in Dortmund. Also, qq plot and residual plots were used to understand whether the model encountered the assumptions of linear model. We did not know about the "No Information" category which included in some parameters but we can presume that especially for No Information category in Energy efficiency class, it had negative influence on price per square meter dependent variable. Goodness of fit was evaluated and score found %30 percent which showed that that proportion of the parameters did explain the output variable.

After making all of these analysis and findings, at this final part of the report, I gave some new ideas and proposals for the possible further investigations. I would prefer to make this regression analysis with more predictors such as whether the neighborhood is close to city center or some public places such as schools or business areas. Also It would be good to make classification of safety of the neighborhood of the properties. Additionally,for larger observations, using the BIC instead of AIC test would be better to make more preferable regression analysis.

Bibliography

Corrie Bartelheimer. Kaggle immobilienscout24 data. 2020. URL https://www.kaggle.com/corrieaar/apartment-rental-offers-in-germany.

David B. Dahl, David Scott, Charles Roosen, Arni Magnusson, and Jonathan Swinton. *xtable: Export Tables to LaTeX or HTML*, 2019. URL https://CRAN.R-project.org/package=xtable. R package version 1.8-4.

Ludwig Fahrmeir, Thomas Kneib, and Stefan Lang. *Regression.* Statistik und ihre Anwendungen. Springer, Berlin [u.a.], 2007. ISBN 978-3-540-33932-8. URL http://gso.gbv.de/DB=2.1/CMD?ACT=SRCHASRT=YOPIKT=1016TRM=ppn+510939260sourceid=fbw$_b$

Aravind Hebbali. *olsrr: Tools for Building OLS Regression Models*, 2020. URL https://CRAN.R-project.org/package=olsrr. R package version 0.5.3.

Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An Introduction to Statistical Learning: with Applications in R.* Springer, 2013. URL https://faculty.marshall.usc.edu/gareth-james/ISL/.

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

David Robinson, Alex Hayes, and Simon Couch. *broom: Convert Statistical Objects into Tidy Tibbles*, 2021. URL https://CRAN.R-project.org/package=broom. R package version 0.7.8.

Gregory R. Warnes, Ben Bolker, Thomas Lumley, Randall C Johnson. Contributions from Randall C. Johnson are Copyright SAIC-Frederick, Inc. Funded by the Intramural Research Program, of the NIH, National Cancer Institute, and Center for Cancer Research under NCI Contract NO1-CO-12400. *gmodels: Various R Programming Tools for Model Fitting*, 2018. URL https://CRAN.R-project.org/package=gmodels. R package version 2.18.1.

Hadley Wickham. *ggplot2: Elegant Graphics for Data Analysis.* Springer-Verlag New York, 2016. ISBN 978-3-319-24277-4. URL https://ggplot2.tidyverse.org.

Hadley Wickham, Romain François, Lionel Henry, and Kirill Müller. *dplyr: A Grammar of Data Manipulation*, 2021. URL https://CRAN.R-project.org/package=dplyr. R package version 1.0.5.