TU Dortmund

Introductory Case Studies

Project III: Regression analysis

Author: Swapnil Srivastava

July 09, 2021

Contents

1 Introduction

In the current era, when people are more keen in buying properties, it is of great interest to understand and study how and which factors affect the price of the property more so that the owner can make the necessary modifications in his/her property and make it more cost effective.

In this report, rental offers for properties on Immobilienscout24 web portal, located in the province of North Rhine-Westphalia (as of February 20, 2020) are taken out of which only Dortmund city data is considered for our analysis. Firstly we perform our descriptive analysis, where we also remove the missing values and introduce our response variable. Then we select our best model using the AIC and BIC model selection criteria and understand the differences between them. We also estimate the best linear model using AIC and understand the coefficients computed. Then, we also compute the confidence intervals and the goodness of fit for our model.

Apart from the Introductory section, the project report deals with additional 4 sections. The Section 2 deals with the description of the data set used in terms of the definitions of the variables and the quality of the data provided. The section 3 deals with the explanation of various statistical terms used for analyzing the data, various model selection criteria for regression analysis. Section 4 deals with the interpretation of the analysis by using the statistical methods (Section 3) on the given data set. The last section i.e. Section 5 contains the summary which deals with all the interpretations and results, also providing an insight for future analysis.

2 Problem statement

2.1 Description of data set and quality

The data set used here is obtained from kaggle.com which comprises data of one of the major real estate web portal in Germany: Immobilienscout24. It contains data of 12118 rental offers for properties located in the province of North Rhine-Westphalia as of February 20, 2020.

It contains 16 variables as follows, **ID**: a unique identification number, **newlyConst**: *categorical variable* - TRUE if property constructed in 2019 or 2020 or else FALSE, **balcony**: *categorical variable* - TRUE if property has balcony else FALSE, **totalRent**:

Total rent generally comprising of base rent, service charges and heating costs, **yearConstructed**: Construction year of property, **noParkSpaces**: Number of parking spaces with the property, **hasKitchen**: *categorical variable* - TRUE if property has kitchen or else FALSE, **livingSpace**: property size in meters, **lift**: *categorical variable* - TRUE if property has lift or else FALSE, **typeOfFlat**: Type of the property, **floor**: The floor of the property, **garden**: *categorical variable* - TRUE if property has garden or else FALSE, **regio2**: The location of city/ municipality, **condition**: *categorical variable* - *good* if the condition is good, *average* is the condition is fine and *NO_Information* if information is not available, **lastRefurbished** : *categorical variable* - *Over5Years* if the property was renovated more than 5 years ago; *Last5Years* if the property was renovated within 5 years and $NO\_Information$ if information is not available, **EnergyEfficiencyClass** : *categorical variable* - $Aplus/A/B/C$ or $D/E/F/G/H$ as per the energy efficiency class of the building and *NO_Information* if information is not available. For our current project, we will be dealing with the data for only Dortmund city.

The data set contains many missing information. We see that while *noParkSpaces* column has the maximum number of missing values i.e. 420, *totalRent* and *typeOfFlat* have 75 and 21 missing values respectively.

2.2 Project objective

The main objective of the project is to perform the *Linear regression* after pre-processing the given data set. Once the data set is cleaned we introduce two new variables namely *groupOfFlat* which is a categorical variable having 4 categories of the typeOfFlat variable and a response variable sqmPrice which computes the rental price per square meter of the property. The basic aim is to find the variables which are impacting the value of rental price (sqmPrice). For this purpose, we find the best *predictors* for it using two types of selection criteria: Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC). After that we estimate this best model predicted using AIC and hence evaluating the goodness of fit of our model.

3 Statistical methods

3.1 Linear Regression model

Here, we aim at modeling the effect of a given set of covariates (or independent, predictor, explanatory variables or regressors) $x_1, x_2, ..., x_k$ on independent variable y (also called as dependent or response variable). One main characteristic of the regression model is that the relationship between the response variable y and the predictor variables is not a deterministic function $f(x_1, x_2, ...x_k)$ of $x_1, x_2, ..., x_k$, instead it shows random errors. This implies that the response y is random. In order to find the possible values of y, we model the **expected** values of x:

$$E(y|x_1, ...x_k) = f(x_1, ...x_k)$$

The variable ε is used to describe the error term in the model. This statistical term represents the random fluctuations, measurement errors or the effect of factors outside of our control. If the function f is a linear function, the model is termed as **Linear regression model.** (L. Fahrmeir et al., 2013)

3.1.1 Simple linear regression model

In simple linear regression model, we model the relationship between one response variable and one predictor variable: $y = \beta_0 + \beta_1 x_1 + \varepsilon$. Here β_0 is the intercept and β_1 represents the coefficient of the covariate. (L. Fahrmeir et al., 2013)

3.1.2 Multiple linear regression model

In cases when the response variable y is dependent on more than one predictor variables, the linear model has the form: $y = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k + \varepsilon$; where β_k is the coefficient of k^{th} covariate. (L. Fahrmeir et al., 2013)

3.1.3 Multivariate linear regression model

It is of the form : $y_i = \beta_0 + \beta_1 x_{i1} + ... + \beta_k x_{ik} + \varepsilon_i$.; where i = 1,.., n, for a continuous variable y and β_k represents the coefficient of k^{th} covariate. The covariates $x_{i1}, x_{i2}, ..., x_{ik}$

can also be continuous, binary or multi-categorical. The unknown parameters $\beta_0 ... \beta_k$ are estimated (represented as $\hat{\beta}_k$) to get:

$$\hat{y_i} = \hat{f}(x_{i1}, x_{i2}, ..., x_{ik}) = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + ... + \hat{\beta}_k x_{ik} + \hat{\varepsilon_i}$$

It can be well represented using **matrix notations** as below:

$$\mathbf{y} = \begin{pmatrix} y_0 \\ \vdots \\ y_i \end{pmatrix} and \, \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_0 \\ \vdots \\ \varepsilon_i \end{pmatrix}$$

$$\mathbf{X} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1k} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{i1} & \cdots & x_{ik} \end{pmatrix} and \, \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_k \end{pmatrix}$$

or, $\mathbf{y} = \boldsymbol{\beta}\mathbf{X} + \boldsymbol{\varepsilon}$ (L. Fahrmeir et al., 2013)

Residuals Residual can be defined as the deviation between the true(y_i) and estimated value($\hat{y_i}$) of the response variable. It is denoted as $\hat{\varepsilon_i}$. Hence :

$$\hat{\varepsilon_i} = y_i - \hat{y_i} = y_i - \mathbf{x_i}'\hat{\boldsymbol{\beta}}.$$
$$or, \; \hat{\boldsymbol{\varepsilon}} = \boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}.$$

The residuals can be demonstrated as the estimates of ε_i. It contains the variations in the data which cannot be explained by the covariates.
Also, i^{th} **standardized residual** can be calculated as:

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

where h_{ii} is the i^{th} diagonal element of the Hat matrix $\boldsymbol{H}$. (A. Rencher, 2008)

Estimating β coefficients: The unknown β coefficients of each covariate are estimated according to the method of least squares. They are determined as the minimizers of the sum of the squared deviation which in turn provides us the value as: $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$
Hence from above equations we can get :

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

where $\boldsymbol{H}$ is the Hat matrix. (L. Fahrmeir et al., 2013)

Coefficient of determination: Coefficient of determination R^2 may be defined as a **goodness-of-fit** measure. It is the proportion of the sum of squared deviations of the dependent variable from its mean that is explained by the regression model. It is computed as:

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

The *closer* the value of R^2 is to *1*, the *smaller* is the residual sum of squares and *better* the fit is to the data. In other words, if $R^2 = 1$, all residuals are zero with perfect fit to the data.

Dummy coding for categorical variables: Since a data set may contain categorical variables as well, we need to model the effect of such multicategorical variable $x \in 1, ..., c$ with c categories. In dummy coding for such variables having c categories, c-1 dummy variables are introduced.These dummy variables can take only values 0 or 1.

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

In order to avoid multiple possible solutions of the model and make the model identifiable, generally one of the possible dummy variable is discarded. (L. Fahrmeir et al., 2013)

Assumptions: Everything about linear regression models holds provided few assumptions hold true. They are: (L. Fahrmeir et al. (2013))

- **Linear Relationship:** Linear regression assumes a linear relationship (either positive or negative) between the explanatory and response variables.
- **Normality of Residuals:** The residuals of our model should be normally distributed.
- **Homoscedasticity of error variances:** It is considered that the variance of ε_i doesn't increase or decrease with one or more covariates x_k i.e. they have relatively constant variance across all predicted y values i.e. $E(\varepsilon_i) = 0 \, or \, Var(\varepsilon_i) = \sigma^2$.

- **No Multi-collinearity:** This states that each independent variable should not depend on another independent variable for its computation.
- The data is independent and identically distributed.
  Using the above assumptions, our equation become like:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + ... + \hat{\beta}_k x_k \, or \, \mathbf{y} = \boldsymbol{\beta}\mathbf{X}$$

3.2 Hypothesis testing

Null hypothesis: It states that none of the covariates are influential while determining the value of the response variable. This means that the β coefficients of all the k covariates are equal to zero. Hence: $H_0 : \hat{\beta}_1 = \hat{\beta}_2 = ... = \hat{\beta}_k = 0$ (L. Fahrmeir et al., 2013)

Alternative hypothesis: Unlike null hypothesis, alternative hypothesis states that the β coefficients of at least one of the k covariates is not equal to zero i.e influential. Hence: $H_1 : \hat{\beta}_j \neq 0;$ for at least one j $\in$ {1,...,k}. (L. Fahrmeir et al., 2013)

Test statistics: We reject or fail to reject our *null* hypothesis using the below computations:

- **Using t values:**

$$t_j = \frac{\hat{\beta}_j}{s\hat{e}_j} \text{ where } \; s\hat{e}_j = \widehat{Var(\hat{\beta}_j)}^{1/2}$$

  $s\hat{e}_j$ can be defined as the estimated standard deviation of $\hat{\beta}_j$. With the above computed value of t_j and considering the level of significance 0.05, we read the corresponding t value from the t-distribution table and compare. If the t-value calculated is greater than the corresponding t-value from t-distribution table, we *reject* the *null* hypothesis or else we *fail* to reject our null hypothesis. (L. Fahrmeir et al., 2013)

- **Using F tests:** Once we've computed the value R^2, as described previously, with df degree of freedom, we can determine the F value as :

$$F = \frac{df}{k} \frac{R^2}{1 - R^2}.$$

Here, it's interesting to note that smaller value of coefficient of determination yields smaller F value, hence its more likely to retain the null hypothesis in such cases while when the coefficient of determination is close to 1 it yields a comparably larger value of F, making it less likely to retain the null hypothesis. (L. Fahrmeir et al., 2013)

- **Using p-value:** It represents the probability of occurrence of the given event which is being taken into account i.e. in our case, the event is that the covariates are non influential while estimating the value of response variable. If the p-value is less than level of significance (as here 0.05), we get a strong evidence to reject the null hypothesis and if not, then we fail to reject the null hypothesis. As it's two-sided test, we compute p-value as:

$$p - value = 2P(t > |t^*|); \text{ where t* is that statistic for each parameter.}$$

Confidence Interval: As mentioned above while estimating the unknown parameters, we have a random variable which can't be estimated completely instead only the β coefficient is deterministic i.e. it can be estimated. Hence confidence interval gives the range of this *unknown but deterministic* parameters.

$$\left[\hat{\beta}_j - t_{df}(1-\alpha/2) \cdot s\hat{e}_j, \hat{\beta}_j + t_{df}(1-\alpha/2) \cdot s\hat{e}_j\right]$$

It is also important to note that if for a particular coefficient the confidence interval contains 0 (i.e. ranges from a negative value to positive), then it does not concretely signify anything about that predictor variable. While on the other hand if it contains only positive values, it means that particular coefficient will have positive effect on the response variable and only negative values will have negative effect. (L. Fahrmeir et al., 2013)

Goodness of fit: As mentioned above, we calculate the coefficient of determination R^2 to determine the goodness of fit for our model. When we compare different models, the usage of this coefficient is limited as it always increases with the addition of new covariates in the model. Hence to avoid this we also take the number of covariates into account and calculate the *Adjusted* R^2 value as:

$$\bar{R}^2 = 1 - \frac{n-1}{n-k}(1-R^2); \quad \text{where k is the number of covariates.}$$

The adjusted R^2 can be termed as the penalty deduction from the original R^2 value for adding more covariates in model. And as described previously, the *closer* the value of R^2 is to *1*, the *smaller* is the residual sum of squares and *better* the fit is to the data. (L. Fahrmeir et al., 2013)

3.3 Information Criteria

In order to select the best model out of all the models, we have many different criteria to judge upon. 2 of such widely used criteria and which will be used in our project as well may be described as below:

3.3.1 Akaike Information Criteria (AIC)

It uses likelihood-based inference in order to compute the AIC values for each model in question. The model which has the lowest AIC value has the best fit out of the remaining models. It is computed as:

$$AIC = -2 \cdot l(\hat{\boldsymbol{\beta}}_k, \hat{\sigma}^2) + 2(|k\,| + 1).$$

Here, $|k\,| + 1$ is the number of parameters and $l(\hat{\boldsymbol{\beta}}_k, \hat{\sigma}^2)$ is the maximum log-likelihood (ML) value. The ML value is determined such that the probability of observing the response variable $(y_1, ...y_n)$ assumes its maximum for $\beta = \hat{\beta}$ and therefore the sample becomes as likely as possible. Specifically in case of linear models with Gaussian errors, we calculate the ML as: $-2 \cdot l(\hat{\boldsymbol{\beta}}_k, \hat{\sigma}^2) = n\ log(\hat{\sigma}^2)$.
Hence for n observations we have:

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

(L. Fahrmeir et al., 2013)

3.3.2 Bayesian Information Criteria(BIC)

The main difference to likelihood-based inference is that here the unknown parameters $\beta_1, ...\beta_k$ are not considered as fixed, deterministic quantities but as random variables with a prior distribution. Hence, BIC penalizes complex models much more than AIC.

It gives the *best* model which is typically more parsimonious than AIC. But on the other hand, while selecting the best model, it behaves similar to AIC, where the lower value of BIC states a better fit model than the rests. (L. Fahrmeir et al., 2013)

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

Similar to AIC, when computing for linear model with Gaussian errors, BIC can be computed as:

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

4 Statistical analysis

All the following measures and plots were created using the software R in version 4.1.0. (R Development Core Team, 2020)

4.1 Preparation of data set

The statistical methods described in Section 3 are implemented on the given data set. As described in Section 2, our data set contains data of 54 city/municipality where the property is located. Out of these, we will only be concentrating on the data belonging to *Dortmund* city for our project. After removing the entries of other cities/municipalities, we look out for missing values. We remove the *noParkSpaces* column (as it has maximum missing values) and the rows which have at least one missing value. Now, the columns which we are left with are the *independent/predictor* variables for our analysis. The *dependent* variable on the other hand is a new variable which we introduce called 'sqmPrice' which is nothing but the *rental price per square meter* and is calculated by the use of 2 independent variables as: $totalRent/livingSpace$. Then, we also introduce another new *independent* variable, 'groupOfFlat' which contains value of the 'typeOfFlat' variable grouped as following 4 categories :

- **apartment**
- **luxurious_artistic_other** comprising the values:'loft', 'maisonette', 'penthouse', 'terraced_flat' and 'other'
- **r_ground_floor** comprising the values 'ground_floor' and 'raised_ground_floor'

- **roof_half_basement** comprising the values 'roof_storey' and 'half_basement'

The column 'ID' is just a unique identification number and does not affect our analysis, so we can remove this column. Since our data set which we will be dealing with, contains only data for *Dortmund* city, hence variable 'regio2' is redundant as it contains value as *Dortmund* only, it is safe to remove this variable as well. Also, as mentioned in Section 3 (Assumptions), the data set should not contain any predictor variables which are dependent on each other. We see that since 'sqmPrice' is calculated using 'totalRent' and 'livingSpace', it is better if we remove the 'livingSpace' and 'totalRent' variables as well. Also since 'groupOfFlat' is obtained from 'typeOfFlat' variable, we can remove 'typeOfFlat' column as well from our data set in order to have no multi-collinearity. Once we are done with above mentioned steps, we get our final data set which has 468 observations and 12 columns. Hence for our analysis we have:

- **Independent variables or Covariates**: newlyConst, balcony, yearConstructed, hasKitchen, lift, floor, garden, condition, lastRefurbished, EnergyEfficiencyClass, groupOfFlat
- **Dependent or Response variable**: sqmPrice

For the response variable, the data set can be summarized as in table 1. We can see from

Table 1: Description of data set grouped by flat type

groupOfFlat	Min	Q1	Median	Mean	Q3	Max
apartment	6.31	9.28	10.35	10.58	11.40	18.38
r_ground_floor	6.48	8.57	10.24	10.30	11.37	18.48
roof_half_basement	7.32	8.82	9.69	10.16	11.06	16.59
luxurious_artistic_other	6.43	10.42	12.48	12.44	13.53	18.98

the table 1, the mean of sqmPrice for all three groups except for luxurious_artistic_other categories is almost equal while that for luxurious_artistic_other is a bit higher. We also see that the range of sqmPrice for three groups except roof_half_basement, lies approximately from 6.3 to 18.5 while that for roof_half_basement is less and varies from 7.32 to 16.59 only. The sqmPrice is maximum for luxurious_artistic_other while minimum for apartment.

4.2 Linear Regression

Null and Alternative Hypothesis: As mentioned in Section 3, our H_0: $\beta_1 = \beta_2 = ...$ $\beta_k = 0$ while H_1: There exists atleast one covariate whose coefficient is $\neq 0$.

4.2.1 Best subset selection

For our report we have taken into account AIC & BIC for the best subset selection criteria. As already mentioned in Section 4.1, we will be performing the regression analysis on our prepossessed data with 12 covariates. The best subset selection is nothing but selecting the model with minimum AIC and BIC value. We obtain the lowest value of AIC as 1933.92 while that of BIC as 1972.99 and the respective models are:

- **AIC**: sqmPrice ∼ newlyConst+balcony+yearConstructed+hasKitchen+lift+floor+ condition+energyEfficiencyClass+groupOfFlat
- **BIC**: sqmPrice ∼ newlyConst+hasKitchen+lift+condition

We can see from above that while AIC gives us the best model having 9 predictor variables, BIC gives us the best model with only 4 predictor variables. While both the models includes variables : newlyConst, hasKitchen, lift and condition; there are few extra like balcony, yearConstructed, floor, energyEfficiencyClass and groupOfFlat which are included in the best subset which is selected using AIC, it is not in BIC model.

4.2.2 Estimation of the best linear model using AIC

The estimation of the best linear model using AIC can be interpreted using table 2.

Table 2: Estimation of the best linear model using AIC

Coefficients	**Estimate**	**Std. Error**	**t value**	**Pr(>|t|)**	**Confidence interval**		**Significant**
					2.5%	7.5%	
(Intercept)	-8.38	8.24	-1.01	0.31	-24.58	7.82	
newlyConstTRUE	1.52	0.53	2.83	**∼0**	0.47	2.57	*
balconyTRUE	-0.41	0.21	-1.90	0.06	-0.84	0.01	
yearConstructed	0.01	∼0	2.38	**0.02**	∼0	0.02	*
hasKitchenTRUE	1.09	0.25	4.38	**∼0**	0.60	1.58	*
liftTRUE	0.77	0.26	2.93	**∼0**	0.25	1.29	*
floor	-0.13	0.07	-1.84	0.06	-0.27	∼0	
conditiongood	1.11	0.25	4.42	**∼0**	0.62	1.62	*
conditionNO_INFORMATION	-0.01	0.28	-0.06	0.95	-0.44	0.42	
energyEfficiencyClassD\E \F \G \H	-0.70	0.36	-1.93	0.053	-1.42	0.01	
energyEfficiencyClassNO_INFORMATION	-0.81	0.30	-2.76	**∼0**	-1.40	-0.23	*
groupOfFlatluxurious_artistic_other	0.27	0.42	0.66	0.51	-0.55	1.09	
groupOfFlatr_ground_floor	-0.85	0.33	-2.52	**0.01**	-1.52	-1.88	*
groupOfFlatroof_half_basement	-0.61	0.33	-1.86	0.06	-1.25	0.03	

We can see that for all the categorical variables, dummy variables (as described in Section 3) have been included. The p-value of this model comes out to be 2.2e-16 which is less than 0.05. Hence it gives us enough evidence to reject our null hypothesis. Also, from table 2 we can see, few coefficients like newlyConstTRUE, yearConstructed,

hasKitchenTRUE, liftTRUE, conditiongood, groupOfFlatluxurious_artistic_other have positive value of estimate while the rest have negative value. This means that these variables with positive estimate have positive effect on the response variable i.e. sqmPrice. The second column i.e. Std. Error measures the mean amount that particular coefficient estimate varies from the actual mean value of our response variable. We can see while its value is less than 0.5 for almost every coefficient, it is almost 0 for yearConstructed. Another important column the p-value measure helps us to reject or fail to reject our null hypothesis. For level of significance $\alpha = 0.05$, the bold p-values are the ones which are <0.05 i.e. are significant enough to reject the null hypothesis. Hence, the coefficients of variables: newlyConst (=TRUE), yearConstructed, hasKitchen (=TRUE), lift (=TRUE), condition (=good), energyEfficiencyClass (=NO_INFORMATION), groupOfFlat (=r_ground_floor) are not equal to 0 and do hold a relationship with the response variable (sqmPrice). This can also be verified using the **Confidence intervals**. For these significant variables determined using the p-value the confidence interval doesn't contain the value 0, hence making the respective coefficient always $\neq 0$. It can also be seen that out of the these variables: newlyConst (=TRUE), yearConstructed, hasKitchen (=TRUE), lift (=TRUE), condition (=good) have their confidence intervals value always positive hence having positive impact on sqmPrice. Similarly, energyEfficiencyClass (=NO_INFORMATION) and groupOfFlat (=r_ground_floor) have negative impact on sqmPrice.

Goodness of fit The value of coefficient of determination (R^2) = 0.283 for the best model determined using AIC. This means that 28.33% of the variance which is found in our response variable i.e. sqmPrice can be explained by our predictor variables. The R^2 obtained here is a bit a low and hence it means that it does not provide a good fit to the data. As mentioned in Section 3, we also determine the value of adjusted R^2 which is dependent on the number of our predictor variables and that comes out to be 0.262 which is also low as they are more close to 0 than they are to 1.

Assumptions Our above model holds true only if all the assumptions mentioned in Section 3 hold true:

- **Normality of residuals:** We can check this using the below QQPlot:

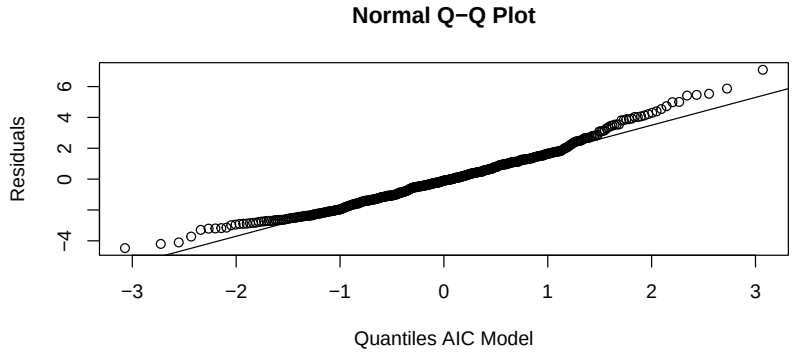

Figure 1: QQ Plot for normality of residuals

It can be seen that since the residuals almost lie on the line, hence normality assumption hold true. Also, it can be noted that since our *lm* function returns value in the R code, this assumption hold true or else it would not.

- It can be seen from figure 2, that the residuals are approximately equally distributed along 0 or else we would observe a cone structure plot instead. Hence, this assumption also holds true.

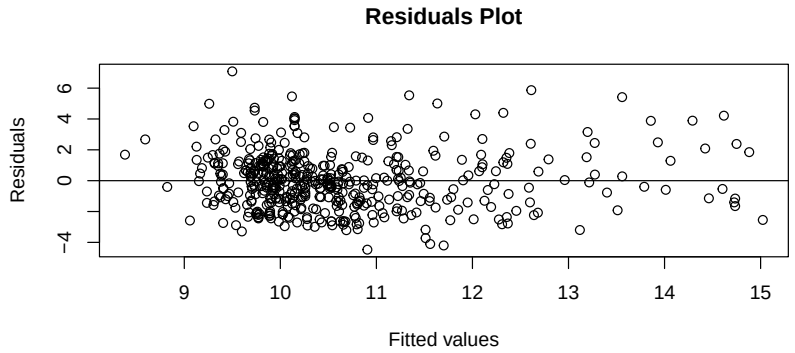

Figure 2: Residual Plot for Homoscedascity of error variance

- The data set is random sample selected from a population, hence the data is independent and identically distributed.
- No multi-collinearity: Since during our data preparation part, we removed all the predictor variables which were dependent on another predictor variable for its computation, our final data set has no multi-collinearity among the independent variables.

5 Summary

This project report deals with linear regression analysis on given data set. The data set contains data of 54 city/municipality of North Rhine-Westphalia. It is collected from kaggle.com which contains data from the real estate web portal of Immobilienscout24. We are only dealing with the data belonging to Dortmund city for this report. Original data set contains 12118 observations and 16 variables but once we remove the missing information and the predictor variables which are dependent on each other, we have 468 observations and 12 independent variables.This also includes two new variables: groupOfFlat (predictor variable) which is a categorical variable having categories of typeOfFlat variable and sqmPrice (response variable) which is the square per meter rent and is computed using two independent variable.

The basic aim of this report is to find the predictor variables which are responsible for increasing or decreasing the rent price for a flat. As nowadays people are more keen in buying/renting flats it is of great interest for the owners to understand and compute the price of their property and also make it more desirable for sale. So once we have our processed data set, we use AIC and BIC model selection which are two different criteria for model selection. It is here that we see that while AIC computes the best model with 9 predictors, BIC best model has only 4 predictors. The p- value of the best model computed using AIC is less than 0.05 hence it gives us enough evidences to reject our null hypothesis stating that there exists at least one predictor coefficient which is not equal to zero. Furthermore using **Confidence intervals**, we also see that out of the 9 predictors computed using AIC, the variabes: *newlyConst (=TRUE), yearConstructed, hasKitchen (=TRUE), lift (=TRUE), condition (=good)* have positive impact on sqmPrice while *energyEfficiencyClass (=NO_INFORMATION) and groupOfFlat (=r_ground_floor)* have negative impact on sqmPrice. We also see the the **goodness of fit** of this model by computing the coefficient of determination which comes to be 0.283 meaning 28.33% of the variance which is found in sqmPrice can be explained by those predictor variables.

Hence to conclude, while few of the variables do not have any impact on the Square per meter rent, most of the variables have. This is due to the preferences and choices of the people willing to buy or rent a property. It may be noted that for this project we only dealt with Dortmund city, for future analysis it would be of great interest to understand and find how the city or region affects the rent price.

Bibliography

G. Schaalje A. Rencher. *Linear Models in Statistics*. John Wiley and Sons, Inc. Publications, New Jersey, 2008.

T. Kneib S. Lang B. Marx L. Fahrmeir et al. *Regression : Models, Methods and Applications*. Springer, Germany, USA, Austria, 2013.

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

Appendix

A Additional tables

Table 3: Data type of variables in given data set

Variable	**Data type**
newlyConst	Boolean
balcony	Boolean
yearConstructed	int
hasKitchen	Boolean
lift	Boolean
floor	int
garden	Boolean
condition	char
lastRefurbish	char
energyEfficiencyClass	char

Table 4: Description of Residual obtained by the best AIC model

Min	Q_1	**Median**	Q_3	**Max**
-4.47	-1.13	-0.09	1.11	7.08