Statistics for Data Analytics - CA1

Multiple Regression Model

Submitted by

Saranya Varshni R K
Student ID - x20154801
School of Computing
MSc. in Data Analytics
National College of Ireland

***Abstract*—Comprehending fundamental features of a property before its purchase is significant so as to make a wise investment decision. Taking such characteristics of house into consideration will improve the chance of picking the best house for the prospective house owners. This document is motivated to identify factors regulating house price and also further focuses on the degree of relationship between the features and the property value. Multiple Regression Model is used to analyze/predict the property price.**

***Index Terms*—Simple Linear regression, Multiple Regression Model, anova, correlation, hypothesis, P- value**

I. Introduction

Understanding the requirements in before hand to the analysis of dataset is the key to obtain a best fit model. With the overwhelming data about n number of properties might push one self into a challenging place to analyse and summarize on the data points incoherent with their housing demands. So, to help in such situations few models shall be considered to provide the best predictions and make a right choice of asset. This document deals with how different attributes of a house influences the property value and also helps in the price prediction of a new house using the multiple regression model. The dataset is divided in the ratio of 80 : 20 as the training and test datasets respectively.

II. Descriptive Statistics and Data Visualization

A. Understanding the features and considering visualization of the same

Categorizing the features in the dataset helps in interpreting the feature type which will further help in variable selection for regression model. The dependent variable 'Price' is a continuous variable that is associated with other independent variables like - lotSize, age, landValue, livingArea, pctCollege, bedrooms, fireplaces, bathrooms, rooms (Continuous Variables) - heating , fuel, sewer (Nominal Categorical Variable) - waterfront, newConstruction, centralAir (Dichotomous Categorical Variable).

Descriptive statistics is the way of organizing and summarizing using numbers and graphs which helps in visualizing data to identify any existing patterns or trends.

Identify applicable funding agency here. If none, delete this.

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation	Variance	Skewness	
	Statistic	Statistic	Statistic	Statistic	Statistic	Statistic	Statistic	Std. Error
price	1728	5000	775000	211966.71	98441.391	9690707465	1.578	.059

Fig. 1. Descriptive statistics on the dependent variable.

Fig. 1. indicates more statistical view that is the test for normality of data by describing the size of the sample that was provided, measure of central tendency (mean, median, mode) and also the variablility of the dependent variable price.

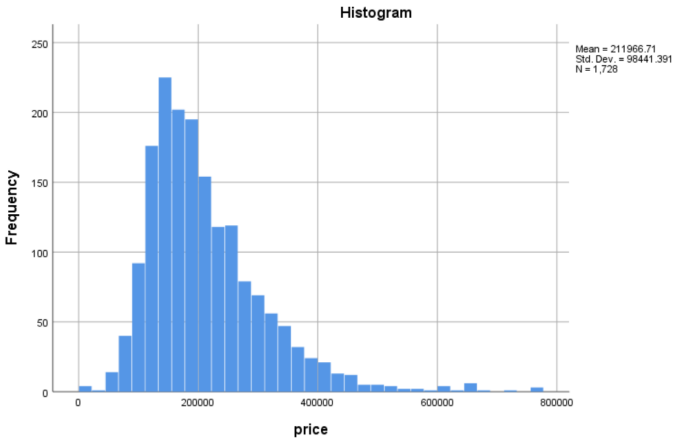

Fig. 2. Histogram of the dependent variable

Fig.2. explains that there is a distribution pattern and is positively skewed. This helps us to understand the spread of the data.

III. Groundwork

Before we begin with regression model there is a lot of data cleaning and preprocessing steps. Dataset wont be clean and neat all the time. In real time streaming data will possess lot of noise that will make the analysis lot more difficult. This might also cause a risk of having the error factor with highest coefficient. So, to reduce that it is necessary to identify the potential outlier. This can be identified using the boxplot.

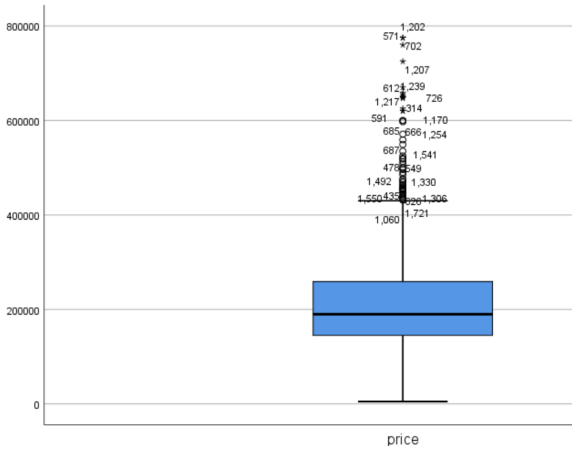

Fig. 3. Boxplot of the dependent variable

It can be observed that this dataset posseses lot of outliers above the wisk in the box plot. Hence, we remove these outliers and we will perform the regression model with two datasets, one is with outlier and another set without the outliers. Once these outliers are removed , the index values of each row is reseeded so as to have a clean data.

If we have one independent variable and one target variable to be tested on the relationship then, Simple Linear Regression would suffice. But with the dataset on house details it is evident that more than one independent variable needs to be considerded in the regression model. Hence, Multiple regression model is utilized here.

In multiple regression the relationship between independent variable and dependent variable are kind of many to one and those independent variables are utilized to explain the variance in the dependent variable or make predictions of the same.

A. *Considerations*

Few considerations that are to be made before performing the regression model is that

- Adding more independent variables to a multiple regression model or procedure does not imply that the regression will be more precise or offer better predictions. In fact it can make the model worse by showing more and more variance in the statistical results. This is termed as OVERFITTING.
- The addition of more independent variable to the regression model creates more relationship among them. So not only are the independent variable potentially related to dependent variable, they are also potentially related to each other. This is coined as MULTICOLLINEARITY. The ideal is for all of the independent variable to be correlated with the dependent variable bi=ut not with each other.

B. *Base analysis - Scatter plot*

- Once the individual data points are plotted on the scatter plot or chart, the relationship or the dependency amoung variables can be identified.

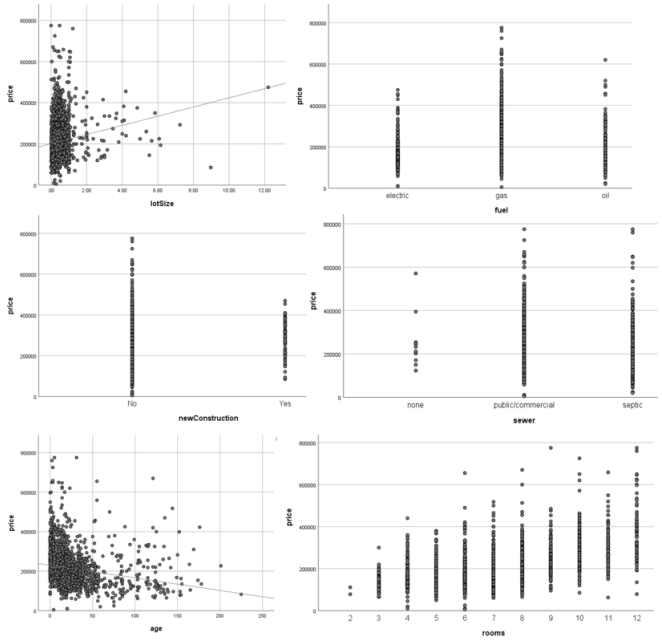

Fig. 4. Scatter plot

- It can be obseved that other independent variable are plotted against 'price'. Considering the lotSize , a linear dependency can be identified where the price increases if the lotSize is bigger. With regards to the newly constructed parameter, major houses are not newly constructed but the dependency with the price is much letter as the new ones have relately moderate price with comparison to old houses.
- In contrast, price of a house decreases with age which is strongly evident with the scatter plot.
- With the number of rooms having quite strong influence on the price as the price increases gradually with the increase in the number of rooms.

C. *Base Analysis - Correlation*

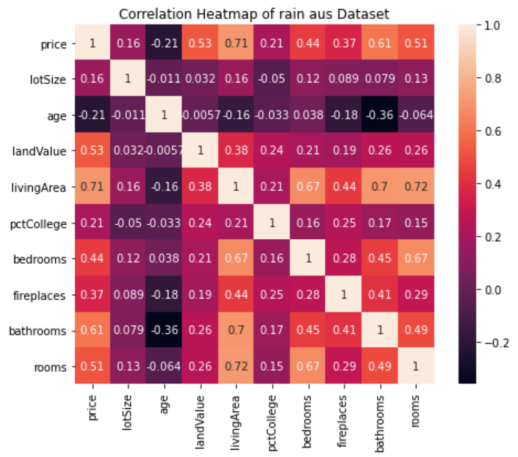

Fig. 5. Correlation heat map

Prior to making selections on the independent variables, to avoid the error of multicollinearity it is necessary to check on the correlation factor. This helps in identifyting the dependency not just between dependent and independent variable but also between independent and independent variables. The base rule is that two independent variables must not be dependent on each other .

Fig.5. represents the correlation heatmap generated using the matplotlib.pyplot library. Here the correlation value varies from 0 to 1 . It can be observed that living area and land value have high degrees of influence on the land price.

D. Simple Regression

In accordance with the correlation heat map , selecting two independent variable that are not much dependent on each other so as to include them in the multiple regression model. the correlation factor shows 0.16 dependency on the living area with age. Hence plotting a scatter plot will help in futher

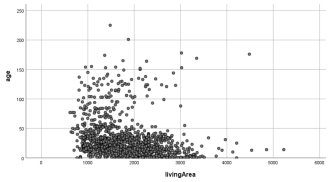

Fig. 6. scatter plot between two independent variables

confirming the relationship. The age independent variable is not dependent on the living area independent variable.

IV. Multiple Regression Model

On consideration of more than one independent variables in the model is called multiple regression model. Here, after performing few simple Linear regression among different independent variables , picking living are and age parameter is most suited and the regression model can be depicted as below

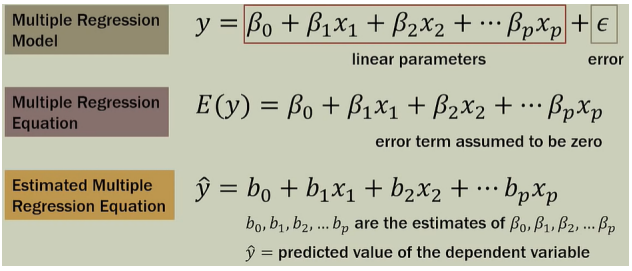

Fig. 7. Multiple regression equation

$$estimatedprice = 111.277LA - 224.751Age + 22971.791 \quad (1)$$

LA= Living area

This model resulted in R square as 0.512 . When the same regression was performed with price as dependendent variable and only living area as independent variable then the R Square was 0.507 where as after the addition another indipendent variable called age resulted in the addition or R Square . Hence the derived multiple regression model is a good model with residual standard error factor as 0.556.

Once the predictor variables are categorical variables then we are suppose to introduce dummy variables. On Generalizing the model we shall use the test data to check on the efficiency of the model that we have concluded with. The same model with the effect of trainging data set tested with the test data set. The below fig depicts the deviations.

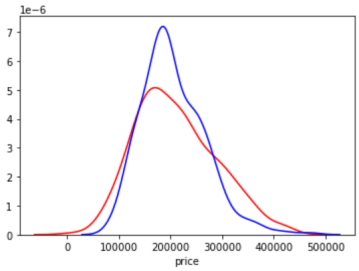

Fig. 8. Difference between training dataset and test dataset for the model created

A. Gauss Markov Assumptions on the model built

In every linear regression or the multiple regression our main aim would be to meet the BLUE which stands for best linear unbiased estimator. This means we want the estimation error variance with minimal value. And this linear regression can be stated as BLUE only if the follow the Gauss Markov assumption.

- According to the 1st assumption in Gauss Markov , we assume that we have our correct functional form of our model.
- The residual variance is constant in the regression model which explains the homoscedasticity assumption. And the same assumption is also made with the model derived

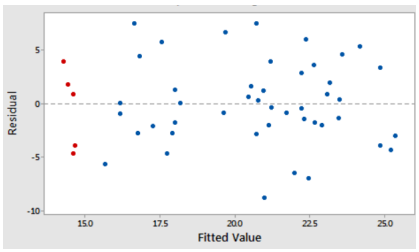

Fig. 9. Residual Vs fitted scale location

- It is assumed that there is no autocorrelation between the errors.
- the Fourth assumption is with the Predictor variables to be independent of the error term taken.

Fig.9. represents the scatter plot between the residual obtained and the fitted scale. Where there is no regular pattern or trend to be observed.

OLS Regression Results

Dep. Variable:	price	**R-squared:**	0.676
Model:	OLS	**Adj. R-squared:**	0.662
Method:	Least Squares	**F-statistic:**	48.54
Date:	Tue, 09 Mar 2021	**Prob (F-statistic):**	0.00
Time:	10:39:30	**Log-Likelihood:**	-20304.
No. Observations:	1674	**AIC:**	4.075e+04
Df Residuals:	1604	**BIC:**	4.113e+04
Df Model:	69		
Covariance Type:	nonrobust		

Fig. 10. OLS Regression regression

V. Summary

To summarize the model built for multiple regression, initially more than one independent parameter was considered for the model. Typically an ideal model for all the individual variables to be correlated with the dependent variable but not with each other. So, with regards to this we have choosen variables that are not dependent with each other but only with the target variable with the help of coorelation factor value , scatter plots and simple linear regression.

It is known that the dependent variable is linearly related to the independent variables, there should be no systematic relationship between he residuals and the predicted (that is, fitted) values. Once the Residuals vs. Fitted graph is plotted there is no fixed pattern or trend is evident.

References

[1] Descriptive Statistics in R —— Exploratory Data Analytics in R —— Data Science
[2] CA1 - Multiple Regression Assignment
[3] Interpret the key results for Descriptive Statistics - Minitab Express
[4] Gauss–Markov theorem - Wikipedia
[5] Homoskedastic Definition