House Price Prediction using Multiple Regression

Adhip Tanwar, Atharv Natekar, Zhaohui Wang, Yushang Chen (GitHub Repository)

This version was compiled on November 14, 2021

Our investigation explores data collected on houses in Saratoga County, New York, USA in 2006. Specifically, it aims to understand which attributes of a house most significantly affect its price. The rationale for this report is to help aid the housing market in estimating the house prices as a function of a multivariate analysis. The report fits the data set into a multiple regression model and conducts a Backward and Forward search using AIC to find out that the house prices were most significantly impacted by the living area and by the inclusion of a waterfront in the property.

House Price Prediction | Regression Analysis | Backward & Forward AIC

Introduction

In the era of globalization, most individuals are focused towards investing their funds in safe assets. There are several objects that are often used for such investments, for example, gold, stocks and property. The property market in particular is one that has observed a constant increase in value over the years. The average house prices in Manhattan for example have steadily increased from approximately USD 500,000 in 2010 to more than a million dollars in 2019

The decision of purchasing a house is based on multiple factors including the age of the house, the location, the size, the neighborhood and various design attributes among the important ones. Since house prices increase every year, there is a need for a system to predict house prices in the future. House price prediction can help the developer determine the selling price of a house. It can also help the buyers to know the price range of the houses, so they can plan their finances well.

This research aims to create a house price prediction model using multiple regression where the house price is considered the dependent variable and all other variables in the data set are considered the independent variables. A backward and forward search using alkaline information criteria (AIC) is used for selection of the most significant variables in this house price prediction model.

Data Set

The given data set - *Housing prices GE19* - focuses on houses in Saratoga County, New York, USA in 2006. It is a random sample of 1734 houses taken from the full Saratoga Housing Data (De Veaux). The data set is sourced from *mosaicData* package in R. The data was originally collected by Candice Corvetti and used in the "Stat 101" case study "How much is a Fireplace Worth". The data set explores 16 variables of interest. *Price* (in US dollars) is our primary dependent variable under consideration. Apart from that, the data set contains 3 categorical variables - *heating* (type of heating system), *fuel* (fuel used for heating), *sewer* (type of sewer system). The remaining numerical variables in the data set include *Lot.Size* (in acres), *Age* (age of house in years), *Land.Value* (in US dollars), *Pct.College* (percent of neighborhood that graduated college), *Bedrooms*, *Living.Area* (in square feet), *Rooms*, *Bathrooms*, *Fireplaces*, *New.Construct* (whether the property is a new construction), *Central.Air* (whether the house has central air), and *Waterfront* (whether property includes waterfront).

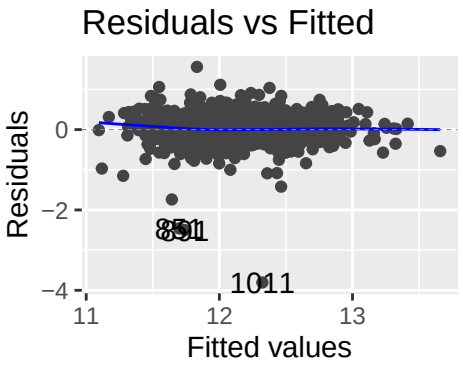

Fig. 1. Residual Plot

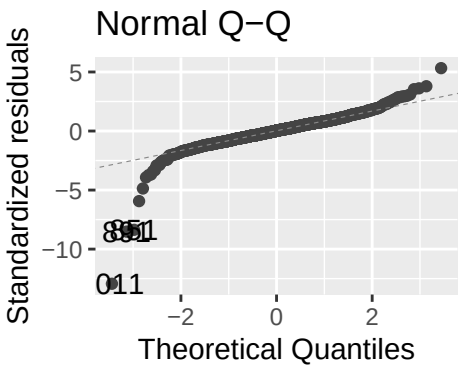

Fig. 2. Q-Q Plot

Analysis

Data Cleaning & Transformations. We are analysing the relationship between the house prices and the other variables in the given data set. Upon initial screening of the data set, we realised that 3 categorical variables - *Heat.Type* (type of heating system), *Fuel.Type* (fuel used for heating), *Sewer.Type* (type of sewer system) - had more than 2 sub-categories. We therefore decided to exclude these variables from our analysis. Upon conducting linearity checks of the variables in our data set with the response variable *Price*, we found that by using the transformation of taking the logarithmic values of *Price*, *Living.Area* and *Land.Value*, we were able to meet the linearity assumption to a greater extent.

Assumption Testing.

- **Homoskedasticity**: In Fig.1 (Residual Plot), no apparent fanning out of the residuals over the range of the fitted values can be observed. The constant error variance assumption is therefore reasonably satisfied
- **Linearity**: In Fig.1 (Residual Plot), when we look at residuals versus fitted values, the blue line is a straight line and does not form any distinct pattern (such as a smiley or frowny face). Linearity assumption can thus be met.
- **Independence**: This sample contains 1734 houses, which are randomly selected from the Saratoga Housing Data (De Veaux)

and therefore, the data set can be considered independent.

- **Normality**: In Fig.2 (Q-Q Plot), most of the points on the QQ plot are relatively close to the diagonal line. Barring a few outliers, the points seem to fit on the line. Therefore, the data seems to be normally distributed.

Model Selection. After ensuring that our data set meets the required assumptions, we proceed towards the final Model Selection.

We first conducted a Backward Search using AIC. Here, we started with a linear model that contains all the potential explanatory variables and performed a backward variable selection to remove variables that are not significant to the data.

We then conducted a Forward Search using AIC. Here, we started with a model containing no explanatory variables. For each variable in turn, we investigated the effect of adding the variable from the current model and only added the variables supplying the most significant information about our response variable *Price* using the add1() function.

In both the Backward and Forward model, R determines the significance of each independent variable by performing F-tests and calculating their p-values.

Finally, we performed a comparison between the model derived using the forward and backward search. We found out that both these models are identical and give us the same significant variables, R-Squared values, and AIC values. We can therefore pick either one of them as our final model.

Hypothesis Testing.

$$H_0 : \beta_0 = \beta_1 = \beta_2 ... \beta_{10} = 0 \; vs \; H_1 : \beta_i \neq 0 \; for \; i \in [1, 10]$$

For formally testing our hypothesis about the relationship between our dependent variable *Price* and our derived independent variables from the final regression model, we set our null hypothesis H0 as all coefficients of the independent variables (from our regression model) equal to 0 VS our alternative hypothesis as at least one of these coefficients not equal to 0.

In order to check our assumptions, we again derived a Residuals Vs Fitted Values plot and a Q-Q plot on our final model and found out that both those plots were identical to Fig.1 and Fig.2 (i.e. the same plots derived on our full model of all variables). We therefore had strong evidence for all our assumptions to be reasonably satisfied (the reasoning explained as above in the Assumption Testing section).

Our Test statistic shows the difference between our full model and our reduced model after dropping the insignificant variables. The observed Test Statistic was **243.2** and P-Value was **< 2.2e-16**.

Since the P-Value was less than 0.05, we rejected the null hypothesis, as we had strong evidence suggesting that there is a significant linear relationship between Price and the derived independent variables.

Results

Final Model.

$$\begin{aligned} log(Price) = {} & 6.772 + 0.0344 * LotSize + \\ & 0.532 * Waterfront - 0.001 * Age + \\ & 0.129 * log(LandValue) - 0.106 * NewConstruct + \\ & 0.060 * CentralAir + 0.530 * log(LivingArea) - \\ & 0.002 * Pct.College + 0.106 * Bathrooms + \\ & 0.0126 * Rooms \end{aligned} \quad [1]$$

- Intercept = 6.772, meaning that considering all attributes of the house to be negligible, the logarithmic value of house price will be 6.772.
- Waterfront has a coefficient of 0.532 which is the biggest coefficient in this model, indicating that Waterfront is the **most significant** variable determining the house price.
- Age has a coefficient of 0.001 which is the smallest coefficient in this model, indicating that Age of the house (in years) is the **least significant** variable determining the house price.

Model Performance. The R-Squared value is 0.585, which means that 58.5% of the variation in the logarithmic value of house prices can be explained by the derived 10 independent variables in our final model (from the previous section).

Model Interpretation. For each of the independent variables and their respective coefficients derived in our final model, interpretations can be made in the following manner:

- On average, holding the other variables constant, a 1 unit increase in Waterfront leads to a 0.532 unit increase in house prices
- On average, holding the other variables constant, a 1 unit increase in Age leads to a 0.001 unit decrease in house prices

(Similar format interpretations can be made about each independent variable in our final model)

Final Discussion

Limitations.

- Some categorical variables such as Heat.Type, Fuel.Type and Sewer.Type had to be removed from our analysis since they had more than 2 sub-categories. The inclusion of those variables could have perhaps resulted in a more accurate house price prediction model.
- Our model only explains 58.5% of the variation in the logarithmic value of house prices. This implies that predictions made using this model have a reasonable scope of error.
- The given data set only contained infrastructural attributes about each individual house. However, house prices also fluctuate heavily by exterior factors such as neighborhood, distance from school zones, floor number, home loan rates etc. In future research, collection and inclusion of data on such external factors could help in forming a far more accurate model for predicting house prices.

Conclusions.

- During our research, we determined that the variables having a significant impact on house prices in Saratoga County, New York in 2006 were Lot Size, Waterfront, Age, Land Value, New Construct, Central Air, Living Area, Pct.College, Bathrooms and Rooms.
- Of these variables, the Waterfront and the Living Area variables affected the house prices most significantly, whereas the Age and Pct.College variables affected the house prices least significantly.
- The model derived in this report only explains 58.5% of the variation in house prices and there is therefore a need to further improve this model, which can be done by including data on multiple external factors affecting house prices.

References

- Andrews, J. (2019). NYC home prices nearly doubled in the 2010s. What do the 2020s hold? Curbed NY. https://ny.curbed.com/2019/12/13/21009872/nyc-home-value-2010s-manhattan-apartments.
- Baptiste Auguie (2017). gridExtra: Miscellaneous Functions for "Grid" Graphics. R package version 2.3. https://CRAN.R-project.org/package=gridExtra
- David Robinson, Alex Hayes and Simon Couch (2021). broom: Convert Statistical Objects into Tidy Tibbles. R package version 0.7.10. https://CRAN.R-project.org/package=broom
- Hadley Wickham, Romain François, Lionel Henry and Kirill Müller (2021). dplyr: A Grammar of Data Manipulation. R package version 1.0.7. https://CRAN.R-project.org/package=dplyr
- H. Wickham. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2016.
- Lüdecke D (2021). *sjPlot: Data Visualization for Statistics in Social Science*. R package version 2.8.9, <URL: https://CRAN.R-project.org/package=sjPlot>.
- Max Kuhn (2021). caret: Classification and Regression Training. R package version 6.0-90. https://CRAN.R-project.org/package=caret
- Masaaki Horikoshi and Yuan Tang (2016). ggfortify: Data Visualization Tools for Statistical Analysis Results. https://CRAN.R-project.org/package=ggfortify
- Randall Pruim, Daniel Kaplan and Nicholas Horton (2021). mosaicData: Project MOSAIC Data Sets. R package version 0.20.2. https://CRAN.R-project.org/package=mosaicData
- Sam Firke (2021). janitor: Simple Tools for Examining and Cleaning Dirty Data. R package version 2.1.0. https://CRAN.R-project.org/package=janitor
- Tarr, G (2021). DATA2002 Data Analytics: Learning from Data. University of Sydney, Sydney Australia.
- Wickham et al., (2019). Welcome to the tidyverse. Journal of Open Source Software, 4(43), 1686, https://doi.org/10.21105/joss.01686