House Prices in New York City

Ishaan Syed Zahiruddin | Jonah Smith | Kunal Ostwal | Luke Gallagher | Yibo Zhao

This version was compiled on November 6, 2022

The goal was to find the possible price prediction for each of the houses. To find features of the dataset that may have a strong correlation with Price, a heat map of the correlation between each pair of variables was generated. The heatmap showed us that Bathrooms, Rooms and Land.Value have a Strong correlation coefficient with Price. Living.Area had the greatest correlation coefficient with Price while Age has a small relationship with Price. With the help of a simple linear model between Living.Area and Price and a multi-variate model where we used both the Forward Stepwise Function(FSF) and Backward Stepwise Function(BSF) which produced the same model. In sample performance showed us that multivariate model has a better goodness of fit compared to the univariate model. 10-fold cross validation test confirmed our findings. From the top three variables with the strongest ability to predict housing price, Price and Living.Area had the highest correlation relationship. This meant that the size of the property increased the price.

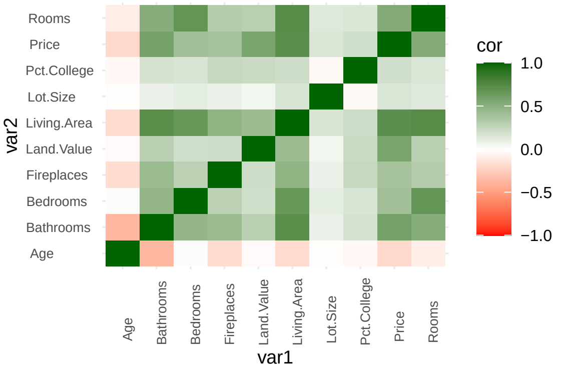
Regression models | Stepwise AIC | NYC Real Estate

Introduction. The USA was hit hard by the COVID pandemic and New York City in particular was the hardest hit city worldwide. The goal of our project was to find the best possible price prediction for each of the houses. The dependent variable was Price of the house. A preliminiary look at the dataset showed us that the New York real estate market is a buyer's market with total sales to listing ratio of 0.12 i.e. the supply of homes is much higher than the demand for homes. Hence, it is a buyer's market.

Dataset Description. The dataset used by us is a random sample of houses from a much larger dataset called the Saratoga Housing Data. The dataset has both categorical and numerical variables which will help to make a more comprehensive prediction. We take it for granted that the data has been independently collected. Since it is a random sample of a larger dataset, the data is assumed to be void of any comfounding factors related to sampling and therefore, we assume there is independence between error terms. Data Cleaning was not an ardous process as the dataset used by us was cleaned by the authors. There were no empty cells. There was an extra variable Test which is not a part of the original Saratoga dataset. Since we do not have a description of it, we omit it from our analysis to remove any comfounding factors from the result.

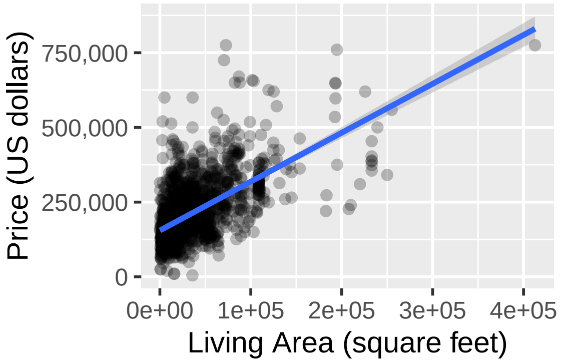
Analysis.

Simple Linear Model. There are a number features of the data set such as `Living.Area`, `Age`, `Land.Value`, that seem like they ought to prima facie have a strong correlation with `Price`. Indeed, a heat map of the correlation between each pair of variables shows a number of interesting things.



The features `Bathrroms`, `Rooms`, and `Land.Value` all have a strong correlation coefficient with `Price`; - surprisingly, `Age` has a small relationship with `Price`; - less suprisingly, `Living.Area` has the greatest correlation coefficient with `Price` of 0.71.

Price vs Living Area



- Consequently, we decided to produce a simple linear model between `Living.Area` and `Price`, described by the equation

$$\text{Price} = 12844.18 + 113.3729 * \text{Living.Area}$$

Stepwise Models. After looking at the result of the simple linear model, we generated a stepwise model to test the dataset's performance in a complex model. We tested both models: the forward and backward AIC function and ended up with the same model.

$$\begin{aligned} Price = {} & 7740.54 \ + \ 70.17 \ * \text{Living.Area} \ + \ 0.92 \ * \text{Land.Value} \\ & + \ 2.304818 \times 10^4 * \text{Bathrooms} + 1.203278 \times 10^5 * \text{Waterfront} \\ & - 4.454452 \times 10^4 * \text{New.Construct} + 9998.56 * \text{Heat.Type(Hot Air)} \\ & - 511.48 * \text{Heat.Type(Hot Water)} \ - 3.295233 \times 10^4 * \text{Heat.Type(None)} \\ & + \ 7372.04 * \text{Lot.Size} - 9639.2 * \text{Central.Air} \\ & - 140.8 * \text{Age} + 3045.91 * \text{Rooms} - 7797.56 * \text{Bedrooms} \end{aligned} \qquad [1]$$

From the generated equation it can be seen that the following factors were not added to the model by the FSF and were excluded by the BSF: `Pct.College`, `Fireplaces`, `Fireplaces`, `Sewer.Type`, and `Fuel.Type`.

Performance Analysis.

In-sample Performance. Here we compare the performance of the multivariate regression model to that of the simple linear regression model within the dataset.

$$\begin{aligned} &\text{Simple Model } r^2 : 0.5090615 \\ &\text{Multivariate Model } r^2 : 0.6548216 \end{aligned} \qquad [2]$$

Hence, the multivariate model has a better goodness of fit compared to the univariate model.

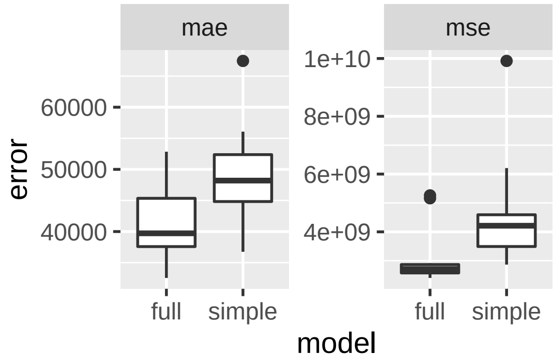
Out of Sample Performance. We chose Mean Absolute Error over Root Mean Square Error since there are several outliers in the data.

We compared the mean absolute error of the two models.

$$\begin{aligned} &\text{MAE(Univariate Model): } 46975.15 \\ &\text{MAE(Multivariate Model): } 41735.76 \end{aligned} \qquad [3]$$

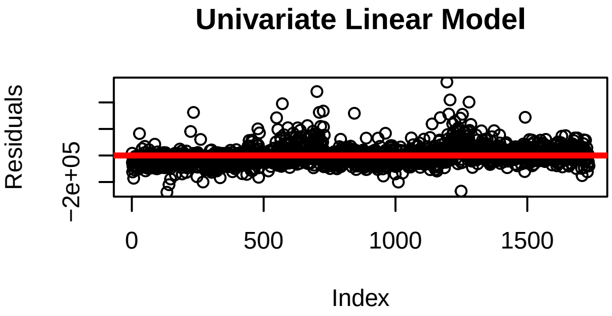
The multivariate model has the lower mean absolute error, as such it is better at predicting the prices of New York houses than the linear model.

10-fold Cross Validation. We performed a 10-fold cross validation test on the two models.



Assumptions.

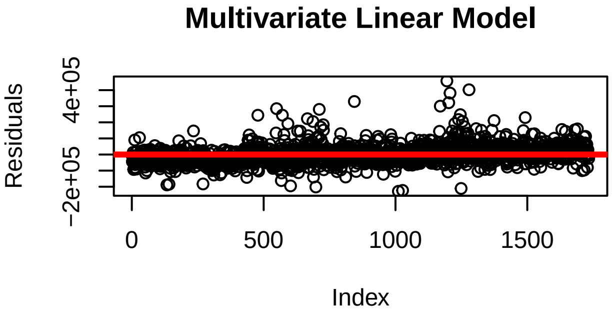
For Univariate model Linearity: The residuals plotted in the univariate model appear symmetrically distributed above and below zero with some outliers above zero, therefore the data is assumed to be linear.



Homoskedacity: In the model,it appears the variance is constant in the residuals plot for each error term and therefore the homoskedacity assumption is satisfied.

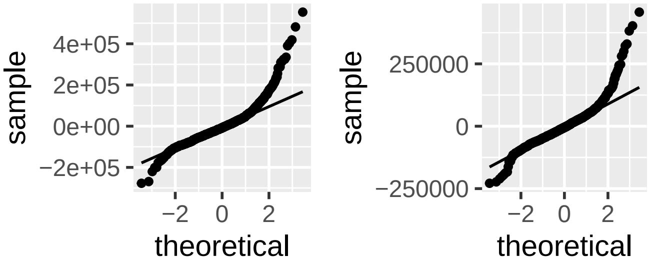
For Multivariate model:

Linearity. The residuals plotted in the multivariate model appear symmetrically distributed above and below zero with some outliers above zero, therefore the data is assumed to be linear.



Homoskedacity. In the univariate model, it appears the variance is constant in the residuals plot for each error term and therefore the homoskedacity assumption is satisfied.

Normality. The majority of points lie close to the diagonal line in both QQ plots, however, there are many outliers in the upper tail so the normality assumption is moderately well satisfied.



Independence. Since it is a random sample of a larger dataset, the data is assumed to be void of any confounding factors related to sampling and therefore, we assume there is independence between error terms.

Results. From our analysis it was found that the top three variables with the strongest ability to predict housing price were Land.Value, Living.Area and Bathrooms. Of these three variables, Price and Living.Area have the highest correlation relationship. When thinking about these results, it seems quite clear that the size of a property would increase the price and the larger the property, the more likely the higher the Land.Value. Bathrooms are an interesting addition, however, it would be fair to say that the larger a house is the more bathrooms it is likely to have. While these seem like obvious predictors of housing prices, these potential assumptions are backed up by the analysis we have undertaken.

In a multiple regression, all factors contributes to the houses price.In answer to our inference: what is set of features that best predicts `Price` we found that:

$$\begin{aligned} Price =& \; 7740.54 \; + \; 70.17 \; * \text{Living.Area} \; + \; 0.92 \; * \text{Land.Value} \\ & + \; 2.304818 \times 10^4 * \text{Bathrooms} + 1.203278 \times 10^5 * \text{Waterfront} \\ & - 4.454452 \times 10^4 * \text{New.Construct} + 9998.56 * \text{Heat.Type(Hot Air)} \\ & - 511.48 * \text{Heat.Type(Hot Water)} \; - 3.295233 \times 10^4 * \text{Heat.Type(None)} \\ & + \; 7372.04 * \text{Lot.Size} - 9639.2 * \text{Central.Air} \\ & - 140.8 * \text{Age} + 3045.91 * \text{Rooms} - 7797.56 * \text{Bedrooms} \end{aligned} \quad [4]$$

Discussion & Conclusions. The increase in housing prices in New York has resulted in many home buyers questioning where and what they can buy, therefore, being able to accurately predict house price based on its attributes is of growing importance. Our analysis led us to the conclusion that there are specific attributes that can help predict housing price, and by using a simple linear model between living.Area and Price we discovered that these two properties had the highest correlation relationship. It would make sense that the size of a home would correlate to an increase in house price, and therefore the need for this prediction may not entirely be necessary. However, in sample performance using a multivariate model, the other variables with the strongest prediction value were highlighted. Bathrooms were found to be a strong predictor of house prices, and showed a trend towards increased price with increased bathroom amount.

While the models are clear and hold in predicting housing prices in New York City there are still limitations to our research. The data used to create the models has a significant number of outliers which have the potential to sway any clear results that may be obtained. Our research also does not investigate trends of specific subgroups of price (such as Fuel.Type) and therefore there may be trends that give rise to situations such as the Simpsons paradox which may decrease the accuracy of the model prediciton.

In summary, our analysis found that the multivariable model generated by the FSF is a good predictor of house prices, and improves the accuracy of any univariate linear model, having satisfied the assumptions required of a linear model.

References.

Github Repository. https://github.sydney.edu.au/kost6112/CC08E2