Analysis of Housing Price Predictors in New York

Luke Gallagher[a], **Jonah Smith**[a], **Kunal Ostwal**[a], **Syed Zahiruddin**[a], **and Yibo Zhao**[a]

[a]DATA2002 Student at The University of Sydney

This version was compiled on November 3, 2022

Using the dataset of housing prices in New York city that identifies the price and the distinguishing features of each home, we hoped to find out whether there is an accurate way to predict the price of a home based on its specific characteristics. It was found through using a combination of simple regression, comparison of a backward and forward model and by comparing the performance of a multivariate regression model to that of the simple regression model that the variables with the highest correlation to housing prices in New York and therefore serve as the best predictors of Price are Living.Area, Land.Value and Bathrooms with Living.Area having the best rate for prediction of overall Price.

Introduction.

- The USA was hit hard by the COVID pandemic and New York City in particular was the hardest hit city worldwide.
- Hence, 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.

What set of factors best predicts price in a linear model?

Data set.

New York Houses.

Source Random sample of houses taken from full Saratoga Housing Data

Variable of Interest

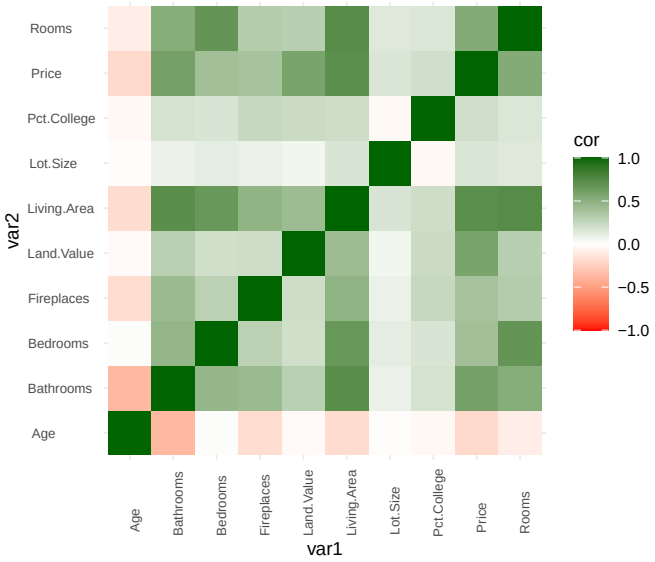
- Dependent Variable: Price of the house
- With this model we intend to find the best prices
- The price is determined by various factors that are included in the dataset

Structure of dataset **price:** price (US dollars)
lotSize: size of lot (acres)
age: age of house (years)
landValue: value of land (US dollars)
livingArea: living are (square feet)
pctCollege: percent of neighborhood that graduated college
bedrooms: number of bedrooms
fireplaces: number of fireplaces
bathrooms: number of bathrooms (half bathrooms have no shower or tub)
rooms: number of rooms
heating: type of heating system
fuel: fuel used for heating
sewer: type of sewer system
waterfront: whether property includes waterfront
newConstruction: whether the property is a new construction
centralAir: whether the house has central air

Data Cleaning

- Since the data is a part of a bigger, more comprehensive dataset, the datasets are already cleaned and nonw of the cells are empty.
- There is an extra variable `Test` which is not a part of the original Saratoga dataset. Since we don't have a description of it, we omit it from our analysis to remove any confounding factors from the result.

Analysis (exploration). 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.

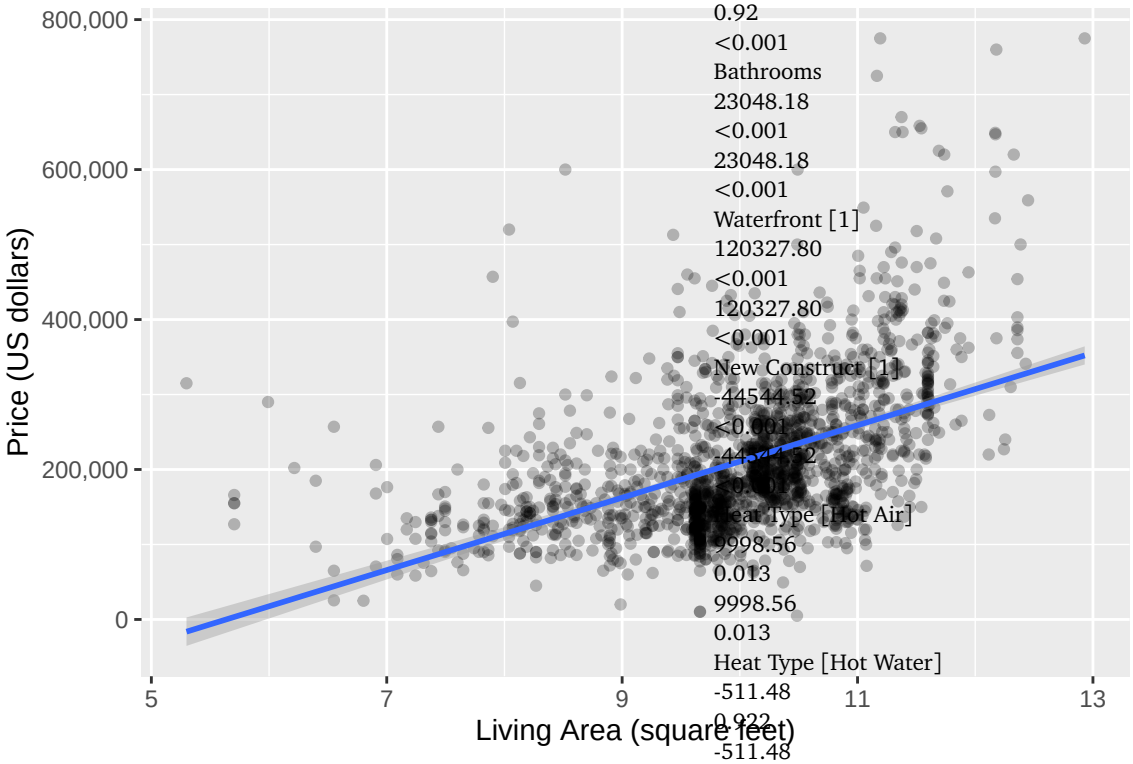
Simple Regression.

Price vs Living Area

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

$$y = 12844.1787 + 113.3729x$$

.



Complex regression.

Stepwise Functions

- In generating a multi-variate linear model, we used the Forward Stepwise Function (FSF) and the Backwards Stepwise Function (BSF). They both produced the same model:

Price ~ Living.Area+Land.Value+Bathrooms+Waterfront+New.Construct+Heat.Type+Lot.Size+Central.Air+Age+Rooms+Bedrooms

- The factors: `Pct.College`, `Fireplaces`, `Sewer.Type`, `Fuel.Type` were not added to the model by the FSF and were excluded by the BSF.

Comparison of backward and forward model.

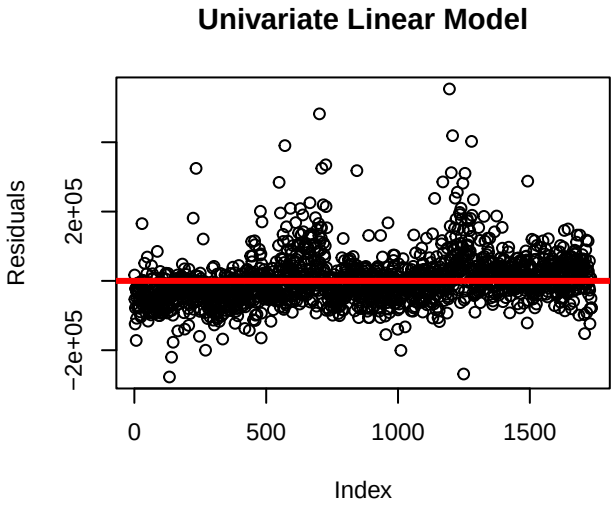
Forward model			Backward model		
Predictors	Estimates	p	Estimates	p	
(Intercept)	7740.54	0.237	7740.54	0.237	
Living Area	70.17	<0.001	70.17	<0.001	
Land Value	0.92	<0.001	0.92	<0.001	
Bathrooms	23048.18	<0.001	23048.18	<0.001	
Waterfront [1]	120327.80	<0.001	120327.80	<0.001	
New Construct [1]	-44544.52	<0.001	-44544.52	<0.001	
Heat Type [Hot Air]	9998.56	0.013	9998.56	0.013	
Heat Type [Hot Water]	-511.48	0.922	-511.48	0.922	
Heat Type [None]	-32952.33	0.182	-32952.33	0.182	
Lot Size	7372.04	<0.001	7372.04	<0.001	
Central Air [1]	9639.20	0.004	9639.20	0.004	
Age	-140.80	0.013	-140.80	0.013	
Rooms	3045.91	0.002	3045.91	0.002	
Bedrooms	-7797.56	0.002	-7797.56	0.002	
Observations	1734		1734		

R2 / R2 adjusted	
0.655 / 0.652	
0.655 / 0.652	
AIC	
42981.739	
42981.739	

Assumptions.

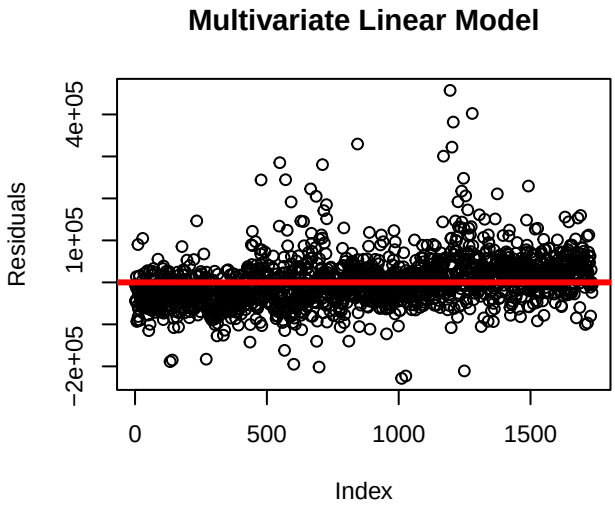
Linearity & Homoskedacity (Univariate)

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



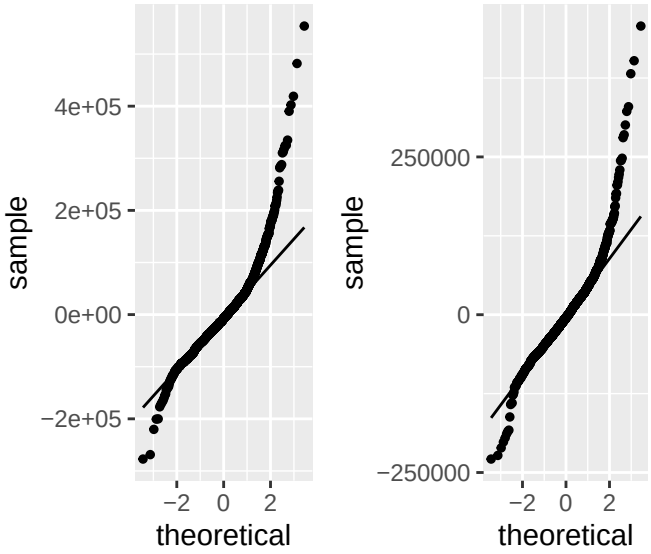
Linearity & Homoskedacity (Multivariate)

- 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 multivariate model too, 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.
- Furthermore, we have a large sample size, and so can use the central limit theorem to justify approximately valid inferences under a normality assumption.



Independence

- 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 confounding factors related to sampling and therefore, we assume there is independence between error terms.

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.

- Simple Model R^2: 0.5090615
- Multivariate Model R^2: 0.6548216

As we can see, the multivariate model has a better goodness of fit compared to the univariate model.

Out of Sample Performance.

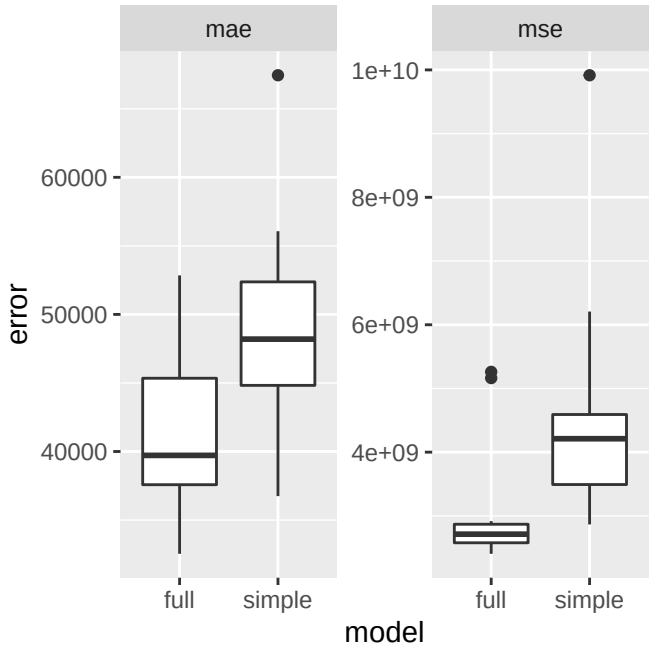
- We chose Mean Absolute Error over Root Mean Square Error since there are several outliers in the data.

We compared the mean absolute error of the two models.

- MAE(Univariate Model) $= 4.697515 \times 10^4$
- MAE(Multivariate Model) $= 4.173576 \times 10^4$

The multivariate model has the lower mean absolute error, as such it has better predictive than the univarite model.

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



Results.

Summary. 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 answer to our inference: what is set of features that best predicts `Price` we found that:

Price ~ Living.Area+Land.Value+Bathrooms+Waterfront+New.Construct+Heat

Provides the best fit of the linear models we tested. It's R^2 indicated a good fit

Discussion and conclusion.

- 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.
- Our analysis did not investigate the trends of subgroups of price (such as `Fuel.Type` and `Energy.Type`. As such, there may be unexplored trends that give rise to situations such as Simposn's paradox. Further limitations include the collection of data

References. Here we differ from PNAS and suggest natbib. References will appear in author-year form. Use `\citet{}`, `\citep{}`, etc as usual.

We default to the `jss.bst` style. To switch to a different bibliography style, please use `biblio-style: style` in the YAML header.

Acknowledgments. This template package builds upon, and extends, the work of the excellent rticles package, and both packages rely on the PNAS LaTeX macros. Both these sources are gratefully acknowledged as this work would not have been possible without them. Our extensions are under the same respective licensing term (GPL-3 and LPPL (>= 1.3)).