Building a Multiple Regression Model on Weight

Zewei Shi(Rio), Calvin Lam, Jia Xu(Sammy), Xuefei He(Faye), Jiao Lu(Cloris)

This version was compiled on November 9, 2020

Pre-assigned a dataset detailing US prices of 1734 New York (NY) households and their respective architectural amenities, this report aims to examine and identify the leading determinants driving NY household prices using a Multiple Linear Regression model whilst simultaneously evaluating the model's appropriateness and performance for this dataset. Leading determinants were selected from amongst the amenities/columns using a backwards-variable selection approach coupled with the Akaike Information Criterion. The resultant regression model found waterfronts, household status' of being newly-constructed, and bathroom counts to be large determinants behind household prices. However, the regressions model itself was found to be questionably-compatible with the data, and thus such observations are likely to be inconclusive.

1. Introduction

The priority aim of this report is to identify the leading household amenity factors provided in the dataset that contribute the greatest to fluctuations in NY household prices, as well as their specific degrees of per-unit influence over property pricing. As a result, the prominent combinations of variables of interest should ideally arise, allowing for generally-optimised decisions of household price budgets contingent on preferred household amenities. This will be achieved via application of a multiple regression model whilst assessing the validity of resultant analyses through model evaluation with respect to the underlying data.

2. Dataset

The dataset text file was imported into a dataframe, specifying its tab-delimited nature. An examination of the dataframe reveals rows of household property prices with column details describing standard household and geographical amenities such as land value and ages of the properties. Further investigation into the source data directed attention towards a dataset named Houses in Saratoga County (2006) with identical column variables along metadata on variable definitions and units of measurements. However, no information could be found regarding data collection methodology. Many of the variables were measured in American units such as price in USD, and lot_size in acres.

3. Analysis

Data Manipulation: Before regressions analysis, it was necessary to explore the data for sufficiently significant variables, noting the immediate removal of the "test" column due to its lack of meaningful purpose. This was initiated with a manual backwards-searching approach to successively remove insignificant amenity/column regressor variables at the 95% significance level. This process was repeated until no variables could be removed of which subsequently, the resultant model was implemented into the Akaike Information Criterion (AIC) function. It was notable that this AIC summary suggested inter-heat_type insignificance; the initial model was thus reevaluated, keeping only the significant variables from the manual back-search approach whilst transforming the heat_type variable into multiple dummy variables with Electric as the base-reference. The aforementioned backwards-search method was repeated on this new model dataset, removing the Hot Water and None heat_type variables to confirm the aforementioned AIC summary. Feeding this new full model back into the AIC function, there existed no more variables with p-values exceeding 0.05, verifying this model as the draft regression model of which model evaluation should follow.

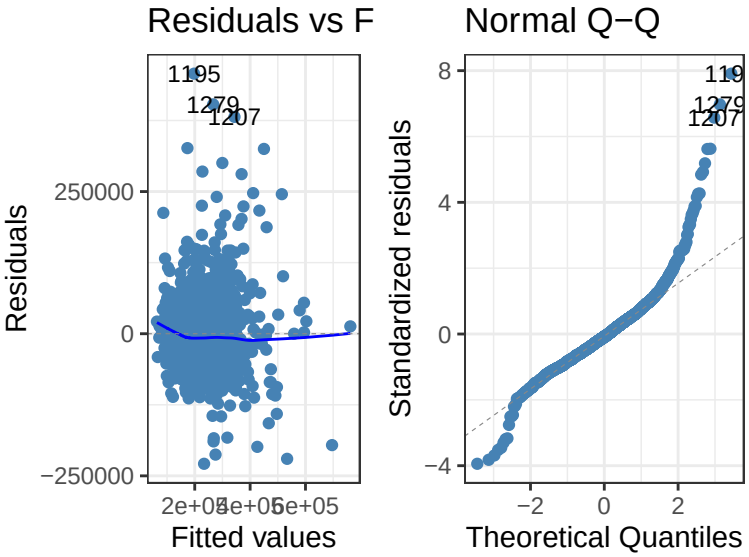

Fig. 1. Residuals Plot and QQ Plot

Assumptions: Checking the distribution of residuals against their hypothetical normal quantiles, there exists a severe violation of residuals' normality and homoscedasticity for only extreme (small and large) values of residuals, otherwise, sufficiently close to the line. When further plotting residuals against the regression model's fitted prices(**Fig.1 Residuals Plot and QQ Plot**), the distribution of residuals around the zero mean was initially inadequate when looking at the entire range of prices. However, when restricting attention to the middle 90% of prices using the "quantile" function from 0.05 to 0.95, the residual's distribution around zero mean appears to be equally-spread around 0. Thus it can be suggested that this regression model is only valid for estimating non-extreme household prices; this also further suggests that the linearity assumption is only met for this range of non-extreme prices. In addition, the independence of observations is indiscernible due to a lack of information regarding exactly how distinct variables were measured, potentially obscuring correlative or dependent relationships between residuals.

4. Results

$$\text{Price} = 7629.14 + 7328.53(\text{lot\_size}) + 120646.11(\text{waterfront}) - 157.3(\text{age}) + 0.92(\text{land\_value}) - 44924.44(\text{new\_construct}) + 9568.78(\text{central\_air}) + 70.29(\text{living\_area}) - 7672.3(\text{bedrooms} \times \text{rooms}_{\text{bed}}) + 22687.1(\text{bathrooms} \times \text{rooms}_{\text{bath}}) + 3076.43(\text{rooms}) + 10476.35(\text{heat\_type\_hotair}) + \epsilon$$

Coefficients: Interpreting the regression model coefficients, the price intercept estimate of 7629.14 is uninterpretable as a house-

hold possessing zero values for rooms, bathrooms, and lot_size (for example) are illogical. With regards to the leading determinants driving NY household USD prices, the variables with the largest coefficients and their dollar interpretations are as follows: the existence of a waterfront predicts an average price increase of $120646.11; a household being a new_construct predicts an average price decrease of $44924.44; a one count increase in bathrooms predicts a price increase of $22687.11. However, there appears to be some concerning figures such as the status of being newly constructed, which seemingly decreases average household price against the logical conclusion of newer infrastructure being superior and more costly. A similar case arises in bedrooms with its predicted average price decrease per bedroom count increase. These are likely to be flaws sourced from the poor applicability of the crafted regression model, however, could also potentially be attributed to factors specific to the real estate industry - expertise that data science students generally won't possess.

Performance: Testing in-sample performance, the r-squared of the model was 0.654, meaning that around 65% of the observed variation in household prices was explained by the model's regressors. The RMSE of approximately $57900 is quite large, almost a third of the median house price of $189700. Further observing the mean absolute error of approximately $41100 which is more resilient to outliers, this is slightly more acceptable, representing the average amount mis-predicted by the model. Testing out-of-sample performance measures via 10-fold cross validation yield similar results.

5. Limitations

The most major limitation of this analysis is the inability for the multiple regressions model to account for extreme values of price. This explains several incompatibility issues with individual regressor variables and regression assumptions of linearity and heteroscedasticity; these can be considerably improved by limiting the price range to non-extreme values, however as a consequence, misappropriate the model. Perhaps a different model independent of linearity may be more effective in characterising determinants driving household prices. The lack of knowledge regarding how the data was obtained and collected may also pose issues to the assumption of independent observations as this obscures possible sources of inter-variable relationships and bias. There's also the assumption that the dataset, sourcing households from Saratoga County only, reflects all household price behaviours throughout NY which is questionable.

6. Discussion and Conclusion

In this analysis, by interpreting the regression model coefficients we identified some large drivers of household prices including the existence of a waterfront, age of household, and counts of bathrooms, although, to an ambiguous degree of validity and success. We hope this study can provide data analysis for those planning to buy a house in New York, allowing them to buy their ideal house base on their needs and budget.

Name	Value
RMSE	57915.63
MAE	41065.27
Median	189700.00

Fig. 2. In Sample Performance

```
Linear Regression

1734 samples
  11 predictor

No pre-processing
Resampling: Cross-Validated (170 fold)
Summary of sample sizes: 1724, 1724, 1725, 1722, 1723, 1723, ...
Resampling results:

  RMSE      Rsquared   MAE
  54464.45  0.7023595  41433.81

Tuning parameter 'intercept' was held constant at a value of TRUE
```

Fig. 3. Out Sample Performance

7. Link to GitHub

https://github.sydney.edu.au/2020-S2C-DATA2002-T11A-Early-2/T11A_early_2/blob/master/draft.Rmd