Predicting house prices given the presence and extent of a variety of common household features

Abstract
This report aims to provide an accurate model of housing prices in Saratoga County, New York. To determine such a model, we constructed a multiple regression model to evaluate the effect that various characteristics of a property have on its sale price. Important steps in using this approach include checking the assumptions of linearity, independence, homoscedasticity and normality, as well as assessing the model's out of sample performance. All assumptions for the data set were met, and upon assessing the predictive performance of the chosen model, it was determined to be an accurate predictor of housing prices in the county.

Introduction
Houses come in a variety of shapes, sizes, and with an assortment of features that differ from house to house. Although there are many commonalities among homes, there are many features that can change both the appeal of the home, as well as the value an individual places on a potential home. In this report, we construct a multiple regression model using a variety of explanatory variables, in order to determine the relationship between house prices and its various attributes and components, as well as to predict house price based upon the presence and absence of these many components.

Data set
The housing prices data set was collected by Candice Corvetti from Williams College Massachusetts for her senior thesis. The data was taken from Saratoga County public records on single family residential home sales. The data consists of 1734 observations and 17 explanatory variables. For each observation, details about the houses are provided by the county's registered providers (RPS). Each sample provides several numerical variables such as house price, lot size, internal square footage and number of bedrooms/bathrooms. There are also several binary variables such as the waterfront status of the property, and character variables such as the fuel and heating methods used on the property. All house prices are valued in the 2002 US dollar.

Results and Analysis:
Meeting assumptions and transformations
In order to perform multiple regression and create a valid model, there are various assumptions that must be met. These include; linearity, independence, homoscedasticity, and normality.

Linearity: To check for linearity we first plotted each predictor variable against the outcome variable Price to determine the existence of an approximately linear relationship (see figure 1). This was true for all continuous (non-categorical) variables. To ascertain the relationships, we then plotted the fitted values against the residuals, which presented a symmetric distribution of the residuals above and below zero for each variable. For the binary and categorical variables, the linearity assumption cannot be applied.

Independence: The assumption of independence for this data set has been met, with each observation being unrelated to the rest, as there are no double-ups for a single property.

Homoscedasticity: Observing the residuals plotted against the fitted values for each (non-categorical) variable, the homoscedasticity assumption is met, as there is a reasonably constant spread of the residuals across the ranges of the fitted values, omitting trivial outliers (see figure 2)

Normality: To check for normality, we plotted the standardised residuals against the theoretical quantities of the model. From the output, we concluded that the assumption had been fulfilled, with the points all lying on or very close to the qq-line, excluding a few outliers (see figure 3).

As all assumptions have been fulfilled, no variable transformations are necessary.

Creating a model

In order to create a model consisting only of significant variables, we used the backwards variable selection method with the F test, removing variables with the least significance. We also performed the backwards selection method using the Akaike Information Criterion (AIC), in order to ensure that the two methods were consistent in selecting significant variables. excluding insignificant variables with P-values greater than 0.05, Both methods selected the same 13 predictor variables which we used to construct this regression model which predicts house prices:

Price = 1.182 + 3.614(lot size) + 7.860(waterfront) - 2.487(age) + 19.946(land value) - 6.247(new construction) + 2.845(central air) + 2.480(heat type air) - 0.099(heat type water) - 1.334(heat type none) + 15.632(living area) -3.063(bedrooms) + 6.952(bathrooms) + 3.179(rooms)

Predictive Performance

To assess the out-of-sample predictive performance of our model, we performed 10-fold cross validation [where the mean absolute error (MAE) of the model's predictions was calculated for each of the 10 iterations. The mean of the 10 MAE values is then obtained in order to assess and compare the model's performance.] We compared the performance of our chosen model, against the performance of the null model which provides a baseline with zero predictor variables, as well as against the full model which includes all original predictor variables. The performance of our constructed model was comparable to that of the full model with respective mean MAE values of 41462.80 and 414648.61. Our constructed model also performed moderately better than the null model which produced a mean MAE value of 73623.42. Although our chosen model performs better than both the full and null models, the mean absolute error can be regarded relatively high, when considering the range of prices in data ($5000 - $775,000). Thus, it is evident that the model can be improved upon, potentially by introducing other variables, to reduce the mean absolute error.

Discussion and conclusion

Our chosen multiple regression model predicts housing prices in Saratoga county, New York based on 13 variables. Our model satisfies all assumptions necessary to conduct linear regression; linearity, homoscedasticity, independence and normality. Our simplified model also provides better predictions than the full model (predicting housing prices using all available predictor variables), whilst using less variables. The predictive performance test reflects however, a lot of room for improvement in the model. The mean absolute error of $41,462.80 in our chosen model, can have a substantial impact on a property's sale price, considering our data's sales range is $5000 to $775,000. Properties at the lower end of this range may be considerably impacted by this error relative to their sale price.

In conclusion, the constructed multiple regression model is a valid predictor of housing prices in Saratoga county, New York, more effective than the current full model, however, can be improved upon by including factors that more accurately predict property values. Further research into housing sales in the county may provide such factors.

References

Corvetti, C (2007). House Price Capitalization of Education by Part Year Residents, Williams College, Williamstown Massachusetts.

Scott, J., 2010. Data Science: A Gentle Introduction. pp.127-147. Available at: <https://jgscott.github.io/STA371H_Spring2018/files/DataScience.pdf>

Appendix

Our group git Repository can be accessed from here:
https://github.sydney.edu.au/lkri7351/M09B_early_2

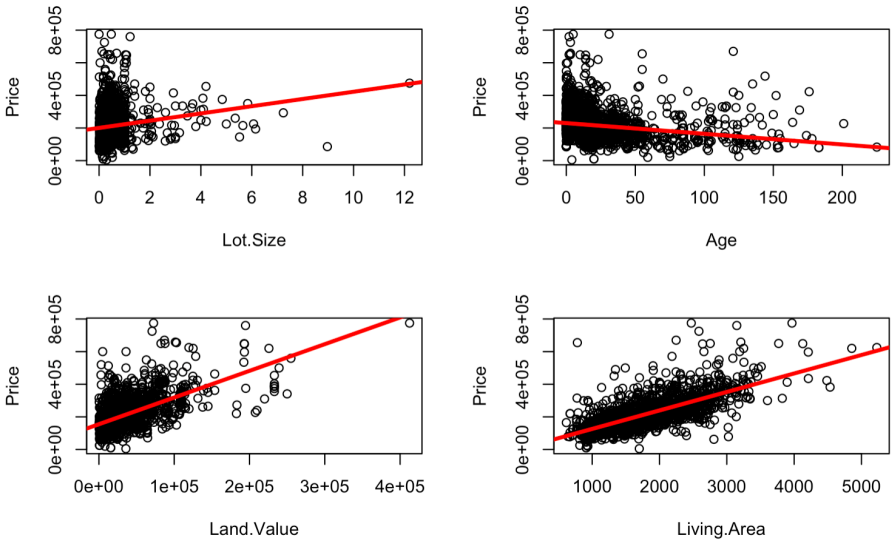

Figure 1: Continuous predictor variables plotted against the outcome variable price, to check for a linear relationship. Assumption is met

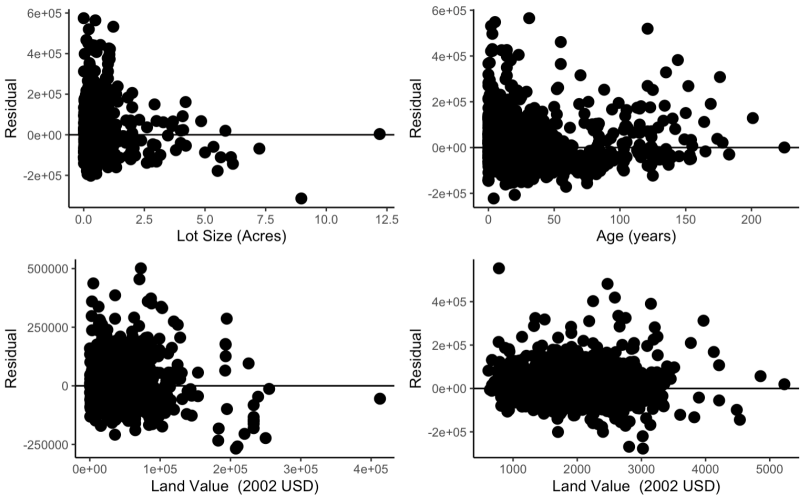

Figure 2: Residuals plotted against fitted values for continuous variables, to check for homoscedasticity. Assumption is fulfilled.

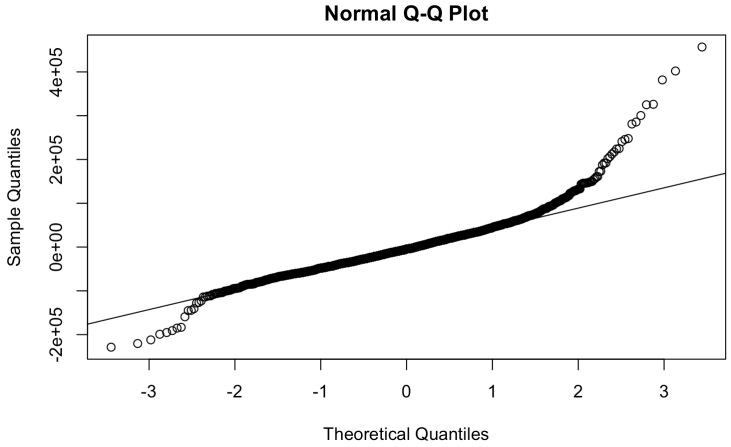

Figure 3: QQ-plot to determine normality. Assumption is met