Multiple Linear Regression Analysis

1. Data Summary

The dataset has information regarding Real Estate in Melborne. It includes factors such as Address, Suburb, Number of Rooms, Landsize, Number of Bathrooms, Price, etc. It has 21 unique variables and 13,580 observations. There are continuous and categorical variables in this dataset. This report will mainly focus on 8 variables namely: ***Rooms*** (int) : Number of Rooms(ranging from 1 to 8), ***Type*** (char): Type of house(h : house; t: townhouse; u : unit), ***Distance*** (double) : Distance from Central Business District(ranging from 0 to 48.1), ***Bathroom*** (double) : Number of Bathrooms (ranging from 1 to 8), ***Landsize*** (double) : Total Land Area(ranging from 0 to 44500), ***BuildingArea*** (double) : Building Size(ranging from 0 to 44,515), ***Regionname*** (char) : General Region(Eastern Metropolitan, Eastern Victoria, Northern Metropolitan, Northern Victoria, South-Eastern Metropolitan, Southern Metropolitan, Western Metropolitan, Western Victoria), ***Price*** (double) : Price of the house(ranging from 131,000 to 9,000,000)

1.1 Objective : There is ought to be a link between the Real Estate factors of Landsize, Building Area, Number of Bathrooms and Bedrooms, Distance from the City Center,etc and the Price of property. The objective of this report is to find how the Price of house is affected by features of the house using Multi Linear Regression. This report also focuses on how the model changes when a subset of features are used to determine the Price.

1.2 Data Cleaning and Preprocessing : This section deals with cleaning and transforming data. The initial step is to drop the redundant data and only keep the variables useful for this analysis. Hence, Rooms, Type, Price, Distance, Bathroom, Landsize, BuildingArea and Regionname variables are selected from the tibble, rest are forwent. The scaled down data is then tested for any missing values, some missing values were found and are removed. The variables Type and Regionname are of type char, but each have 3 and 8 values respectively, these are converted to factor with levels 3 and 8.

2. Planning

The first objective is to check for assumptions about dataset. The assumptions being tested in this section are quantitative or categorical predictors, quantitative and continuous outcome, non-zero variance and predictors are unrelated to external factors.

2.1 Check predictor and outcome variable categories : For Linear Regression, the predictor variables are supposed to be quantitative or categorical. There are 7 predictors in this analysis and are of type factor, integer and double. Hence, all the predictors satisfy the assumption. The outcome variable(Price) is continuous and quantitative, satisfying the assumption.

2.2 Non-Zero variance and predictors unrelated to external factors : The variables for Linear Regression are assumed to have a non-zero variance, after looking at parameter distribution and scatterplots, it is inferred that the parameters hold the assumption of non-zero variance. For the scope of this report, the predictor variables are assumed to be unrelated to external factors.

3. Analysis

This section incorporates performing multiple linear regression on two different models with different predictor variables and comparing the results.

3.1 Models: The predictors used in Model A are (Rooms, Type, Distance, Bathroom, Landsize, BuildingArea, Regionname) and in Model B (Rooms, Type, Distance, Bathroom, Landsize, BuildingArea). The variables Type and Regionname are factors and are converted to contrasts to remove redundancy in the coefficients of multiple linear regression. After applying the models, the residuals are found to be non-normal, heteroscedastic. To get rid of this issue, this report includes the log transformation of the outcome(Price) variable. As, the Price variable is positively skewed, applying log transformation made the Price normally distributed.

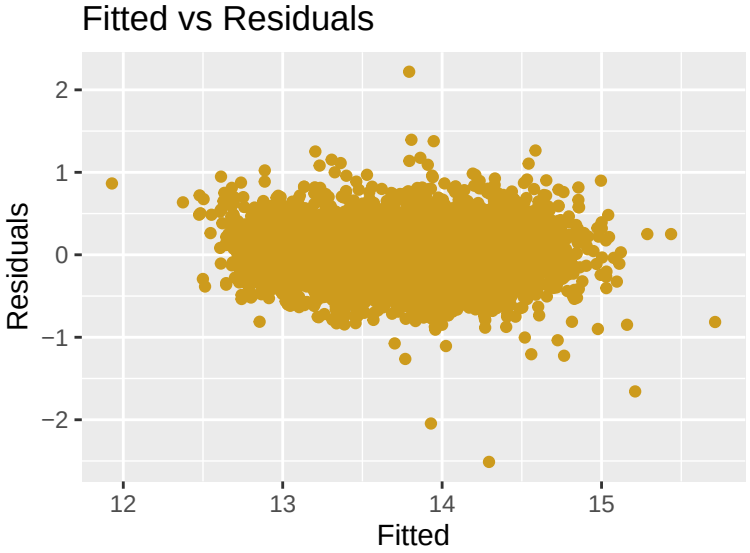
3.1.1 Diagnostics on Models : This section includes performing Outlier and Influential Points testing and remove if any. To test for ***Outliers***, the residuals are standardized and check the data points below -1.96 and above 1.96 standard deviations, i.e the datapoints out of 1.96 standard deviations. About 5% of the datapoints were found to be outside this range in both the models. 95% of standardized residuals are between -1.96 and 1.96. So, the residuals are normally distributed and this report does not consider any Outliers. ***Influential Points*** can be tested using Cook's Distance[1]. One data point in each model has Cook's Distance > 1 stating the presence of an influential point and is removed.

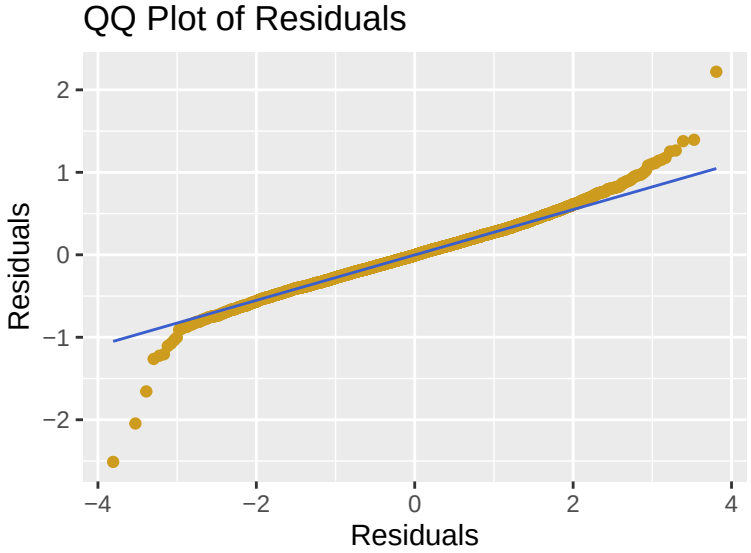
[1] A test to check for Influential Points. If Cooks's Distance > 1, then the point is Influential

3.1.2 Generalization and Assumptions on Models : This section checks the Residuals for homoscedasticity, normality, independence and for linearity and no multicollinearity. ***Independence of Errors*** is tested using a test called Durbin-Watson[2]. For Model A the d value ~ 1.55 at p value = 0, and for Model B the d value is ~ 1.27 at p value = 0, although the d value is different from 2, but this would not break the model and this report proceeds with the assumption of Independence. ***Homoscedasticity of Residuals*** is tested by looking at the scatterplot for residuals vs fitted values, the plot shows the residuals to be homoscedastic thus fulfilling the assumption. ***Normality of Residuals*** is tested by plotting the Q-Q plot, the plot seems to be normal thus this assumption is held true. ***Linearity*** is tested by plotting residuals vs fitted. The plot shows linear relationship thus holding the assumption. ***No Multicollinearity*** is tested by using VIF (Variance Inflation Factor)[3], the VIF for all the predictors are less than zero. The data holds the condition for No Multicollinearity.

3.1.3 Multiple Linear Regression on Model A: There are 15 Coefficients for Model A and 2 out of 15 have p value > 0.05, rendering them non-significant and are not included in the final model. The calculated multiple R^2 for model A is **71%**, i.e. the model explains **71%** of the variance with Adjusted R^2 as **70.94%**. The confidence intervals for all the significant coefficients does not overlap zero, thus the results are significant and with 95% confidence the coefficients lies between those ranges.

3.1.4 Multiple Linear Regression on Model B: There are 8 Coefficients for Model B and none of them have p value > 0.05. The calculated multiple R^2 for model B is **57.1%**, i.e. the model explains **57.1%** of the variance with Adjusted R^2 as **57.06%** . The confidence intervals for all the significant coefficients does not overlap zero, thus the results are significant and with 95% confidence the coefficients lies between those ranges.





3.2 Comparing the models

The best method to compare the two linear regression methods is ANOVA[4]. The results of anova show the p value < 0.05, which indicates that the more complex model is the best fit. In the two models, Model A is complex with 7 predictor variables and Model B with 6 predictor variables. The inference from ANOVA is that the Model A is better representative of the data than Model B.

4. Conclusion

After scaling down the data, checking the assumptions for variable categories and non-zero variance. It is examined that the dataset is eligible for linear regression. The Linear Regression is applied to the model but it generated non normal and heteroscedastic residuals. Log transformation is applied to the outcome variable, this improved the fit of the model and produced normal and homoscedastic residuals. The model is then tested for diagnostics and generalization assumptions of No Multicollinearity, Linearity, Normality & Independence & Homoscedasticity of Residuals. All these condition were held true, stating that the Multiple Linear Regression can be applied to the dataset. The complex Model A performed better in terms of R^2 valued at **71%** whereas for Model B at **57.1%**. These models were further compared using ANOVA to distinguish on statistical level. The results of ANOVA were in coherence with R^2, rendering the Model A to be better at explaining the variance. In nutshell, the complex model performed better and the factors could explain 71% of the variance of the data.

[2]Test used to check the Independence of Errors. D value close to 2 means the residuals are Independent and no autocorrelation is present.
[3]Vif < 10, holds the No Multicollinearity Assumption
[4]Analysis of variance. It tests if the means of the two models is different