Predicting New York house prices using K Fold Cross Validation and Backward Selection to build a Model

# Abstract

The project is designed to identify a model that best predicts price of New York houses based on the 10 factors (area, bathrooms, stories, mainroad, basement, hot water heating, air conditioning, parking, prefarea, furnishing status). K fold Cross validation was used to compare the full model and best subset model of 11 predictors. Results showed that the 11 predictors model (price~ area + bathrooms + stories + mainroad + guestroom + basement + hotwaterheating + airconditioning + parking + prefarea + furnishingstatus) had the lowest

RMSE (1077939) and MAE (789374.1.2).

# Introduction

Housing New York can be expensive depending how far you live from the main city and other factors such as parking, no of bedrooms can increase the costs. This model we are trying to create is help users to predict the cost of their house. So, they can use this information to negotiate a better deal from the real instate company. Having the knowledge of what the price of the house before hand can save customers from overpaying the house they are purchasing.

# Methods

All of the data used in the project is from <https://www.kaggle.com/datasets/yasserh/housing-prices-dataset?resource=download> kaggle a reliable source. Data was clean but required to be label encoded, then scatter plots were created to visually inspect the data. Alpha = 0.05 was used for this project.

|  |  |
| --- | --- |
| Column | Description |
| Price | price of the house the dependent variable |
| area | area of the house |
| bedroom | number of bedrooms |
| stories | floor height of the house |
| mainroad | whether the house is located near the main room |
| guestoom | does the house include a guest room |
| basement | is there a basement in the house |
| hotwaterheating | does the house have warm water |
| airconditoning | is there a ac in the house |
| parking | number of parking allocated to the house |
| preface | Boolean datatype |
| furnishingstatus | status of the furnishing of the house has 3 states |

Table 01: column names and their respective description

Simple Scatterplot Matrix

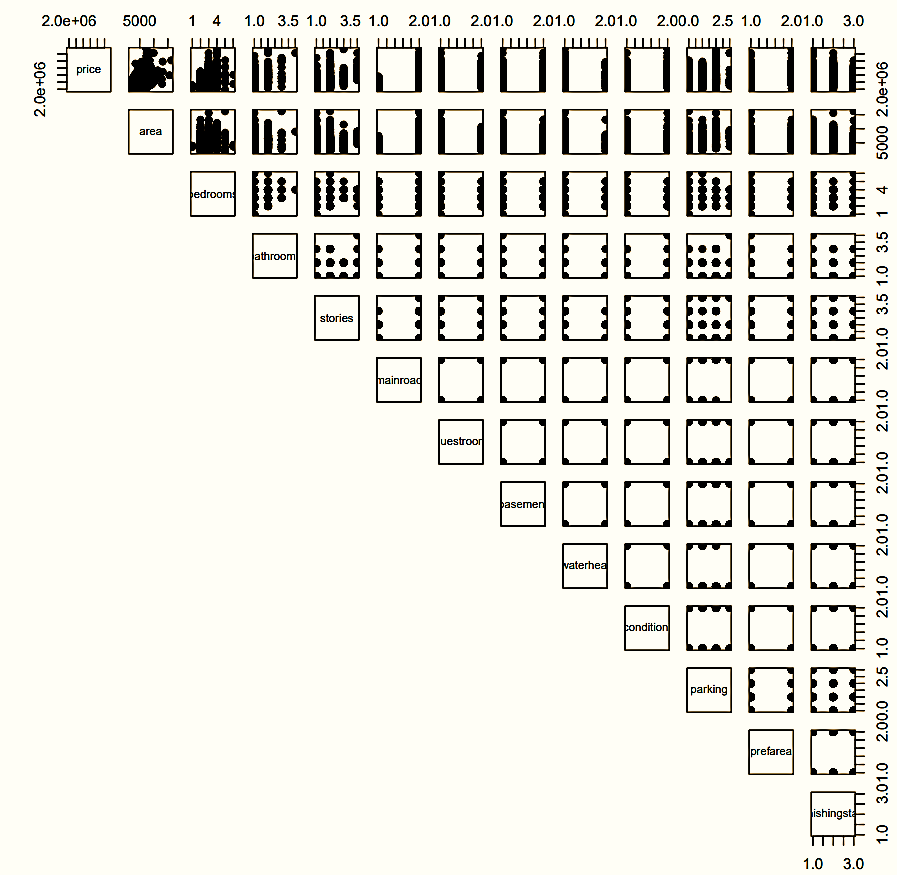


Figure 1.0

There seems to be a multicollinearity issue so we can do a VIF test. However, all the numbers were significantly below ten. Hence there’s no multicollinearity problem. Later a backward stepwise and a both step regression functions were used. Both these functions didn’t result in the same output. Hence, we had 2 models, the full model, and the best subset of 11 parameters.

Subsequently K fold cross validation was used to examine which of the resulting models better predicted house prices. Results from the K fold cross validation are presented in below

# Results

After conducting the K fold cross validation test with the full model and 11 predictor model, we can conclude that the 11 predictor is better as it had the lowest RMSE (1077939) and MAE (789374.1.2).

# Discussion

One of the concerns about this data would be the number of categorical variables present. Since most of these variables were categorical into 2 types deriving a correlation to the predicted variable (price) would be an inefficient method. If we could somehow combine these columns or have more distinct states, we could achieve a more accurate model.

Below this you can observe the Q-Q plot, Residual vs fitted plot, Outlier and Leverage for price and actual vs fitted plots. We can note that in the Q-Q plot the points are closely aligned to the slope just not around the upper quartile. This shows that the distribution is slightly skewed to the right. In the residual plot the average residuals is consistent throughout the plot. About the leverage of points, we have many outliers present in the data which can’t be ignored and should be dealt with. Plotting the regression line of actual and fitted side by side we have a visual perspective of how close the lines are. We can predict best results for houses of low price than higher price.

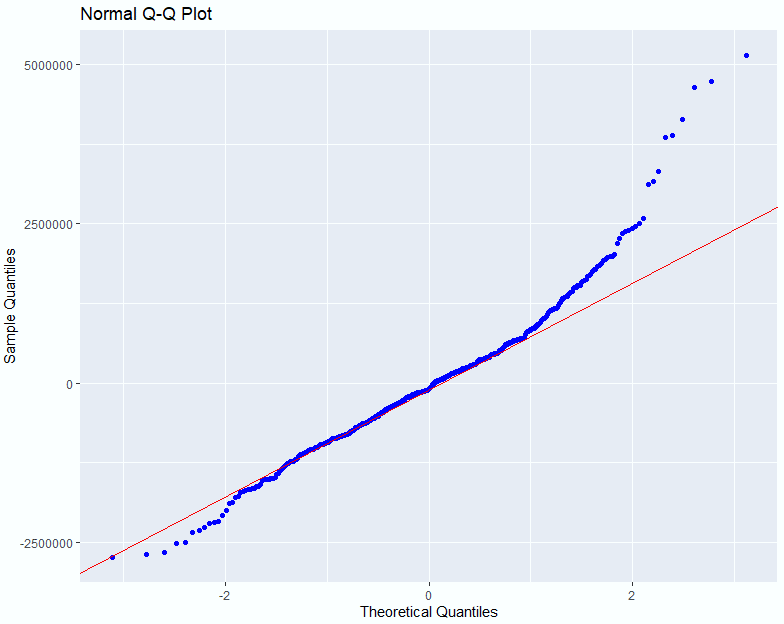


Figure 2.0

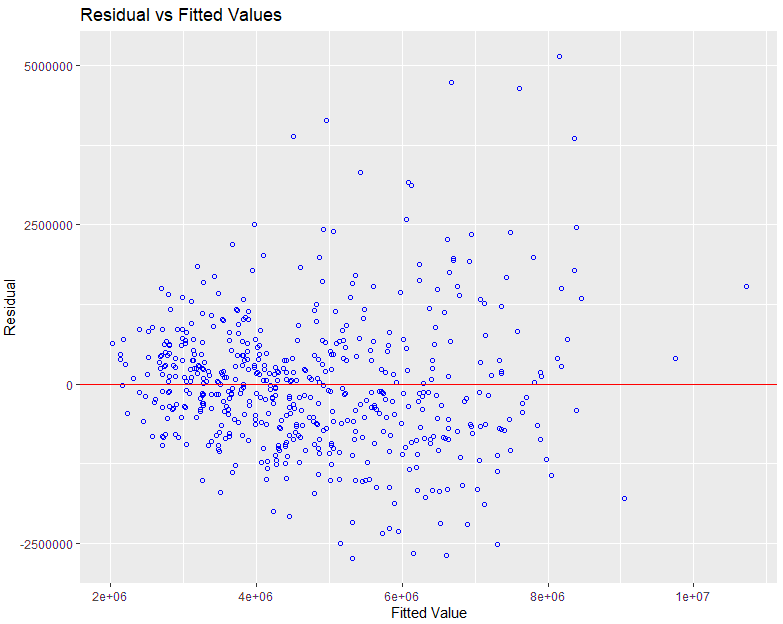


Figure 2.1

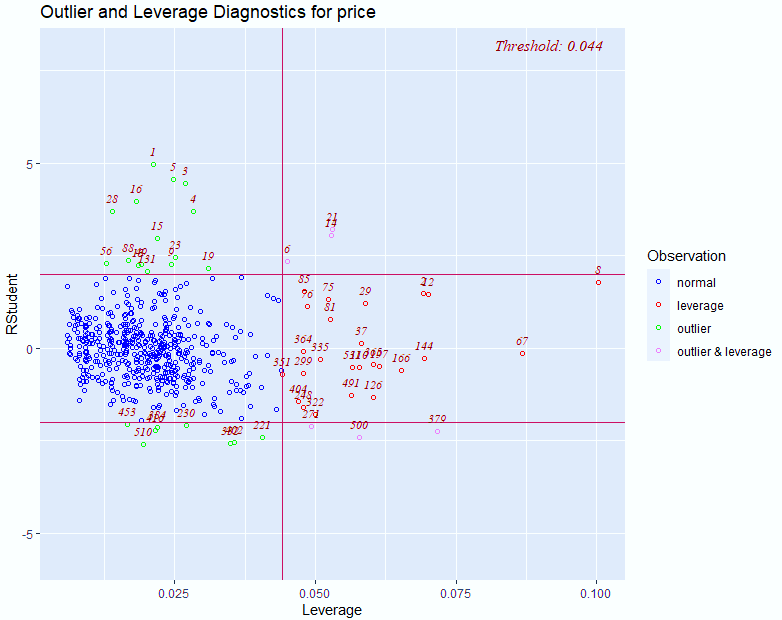


Figure 2.2

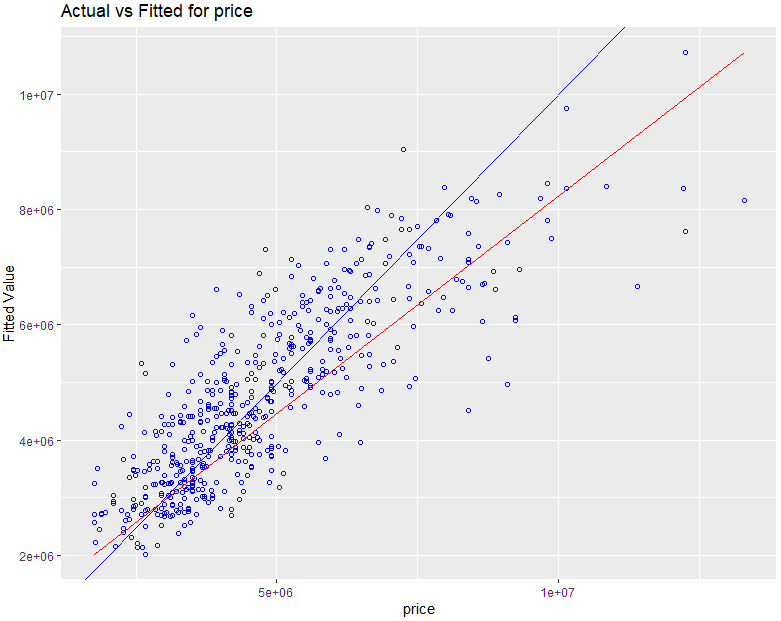


Figure 2.3



Figure 2.4: Test Correlation

# Appendix A

# Output

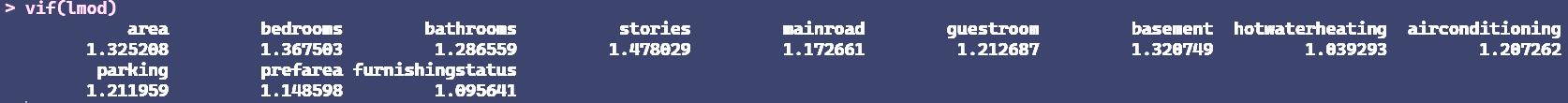


Figure 3.0: VIF of the model

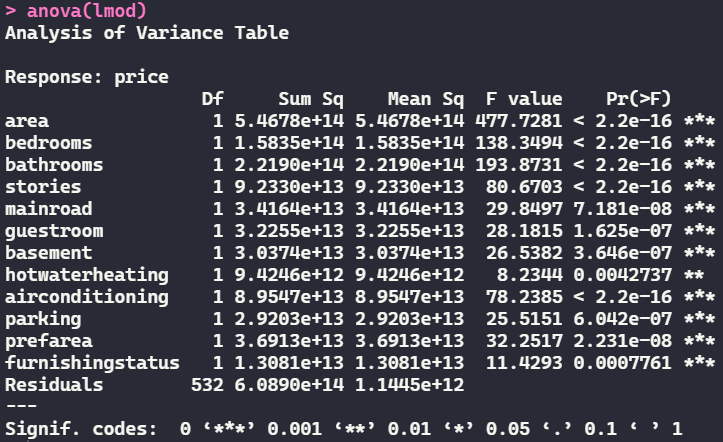


Figure 3.1: Anova of the full model

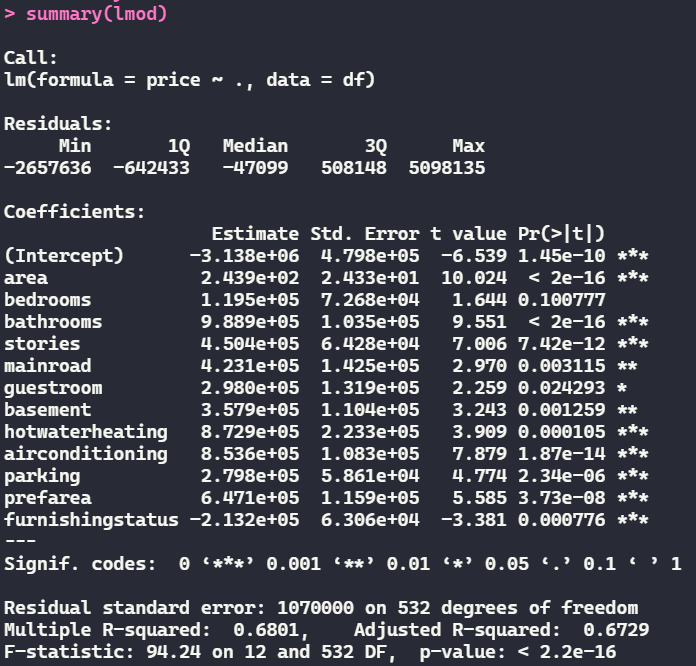
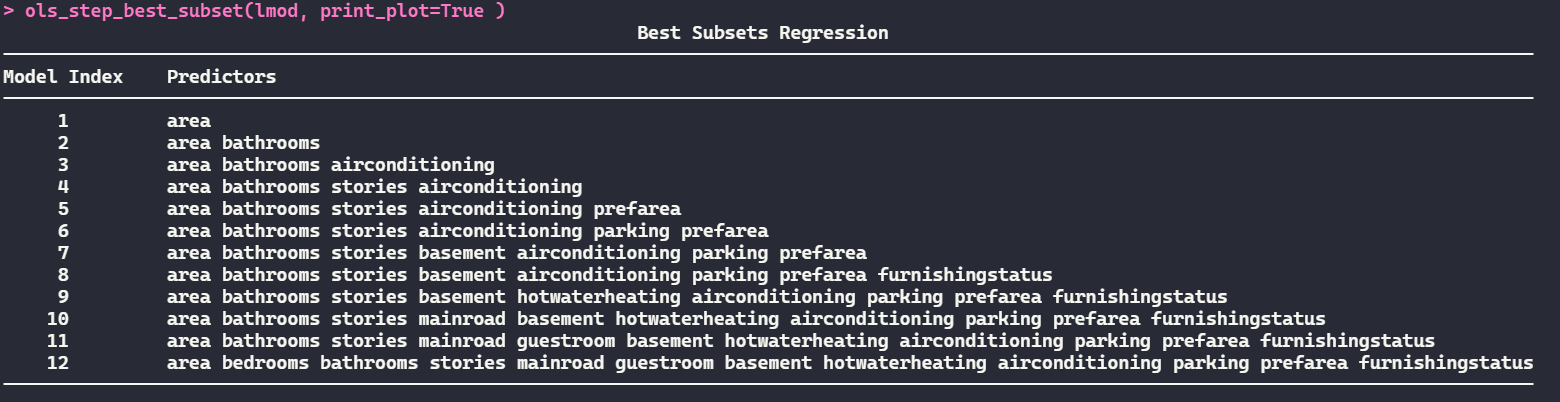


Figure 3.2: Summary of the full model



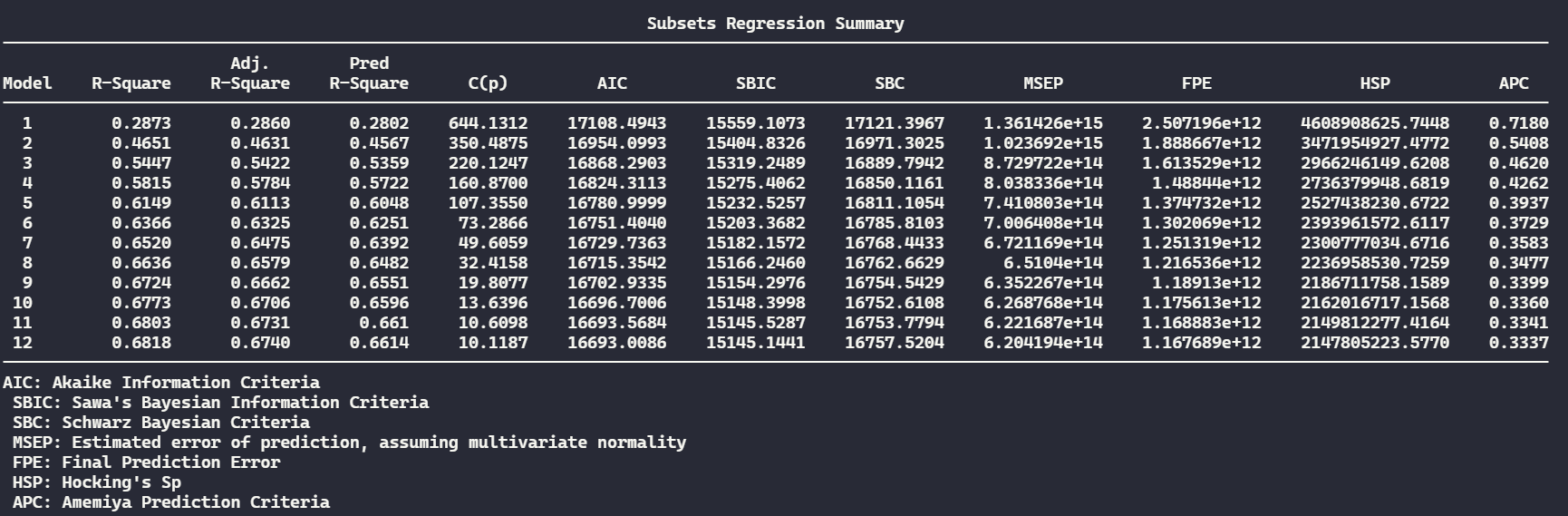


Figure 3.3: Best subset for the model from 1 predictor to 12 predictors

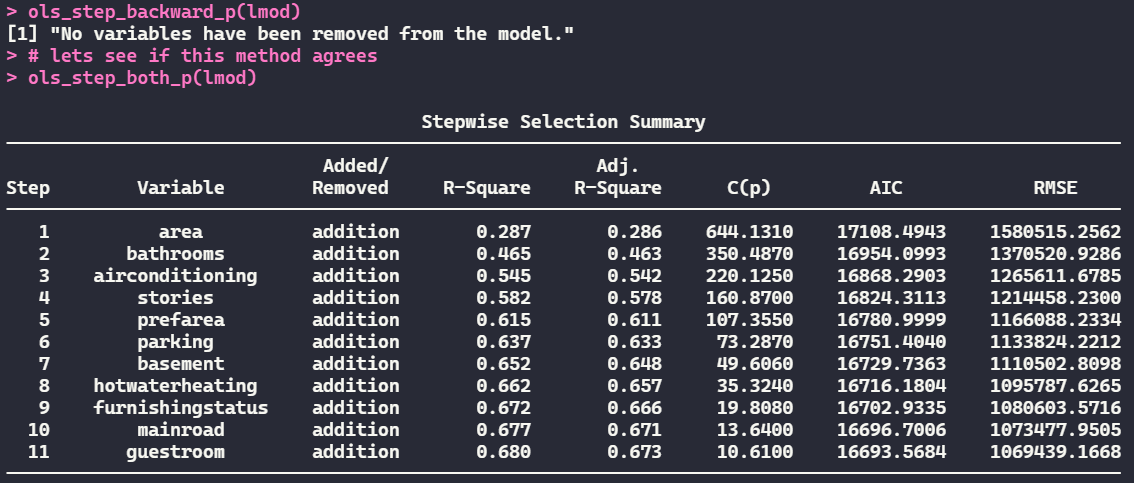


Figure 3.4: Backward and Selection and step both ways methods

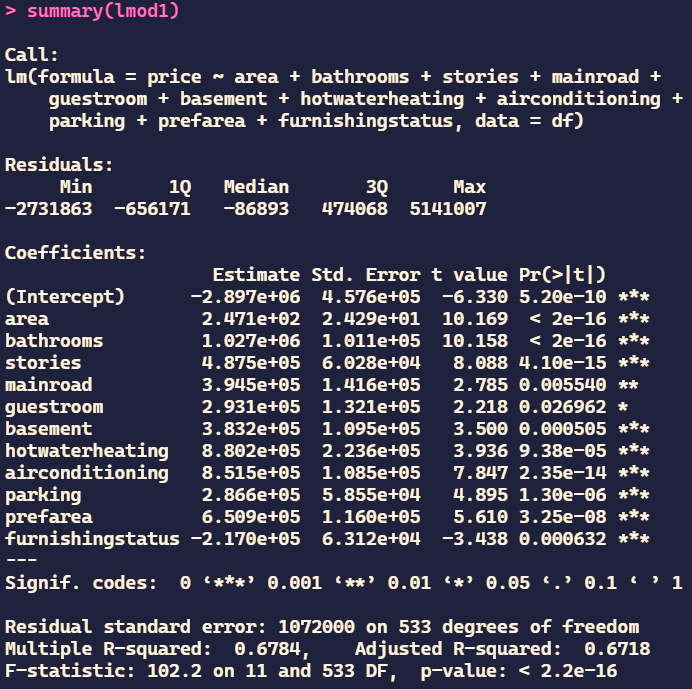


Figure 3.5: Summary of the model with 11 predictors

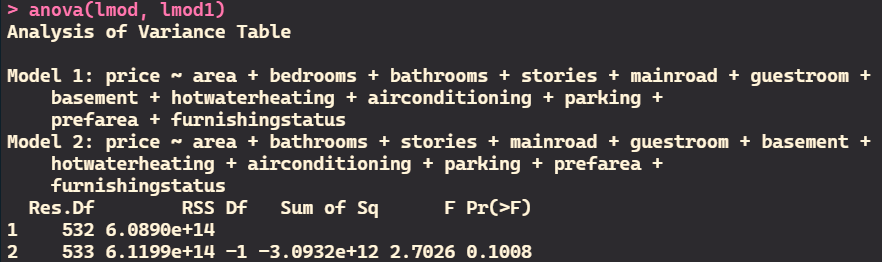
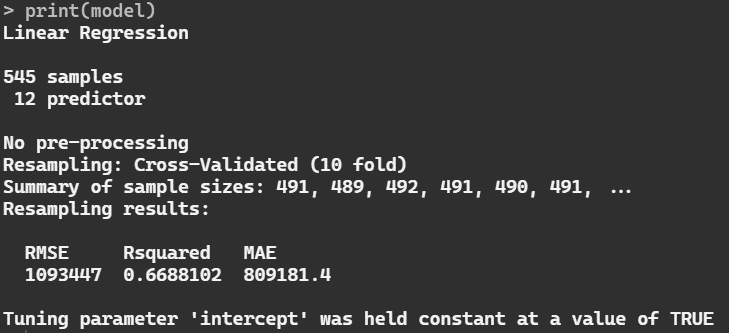


Figure 3.6: Anova test to compare significance of full model and 11 predictors model



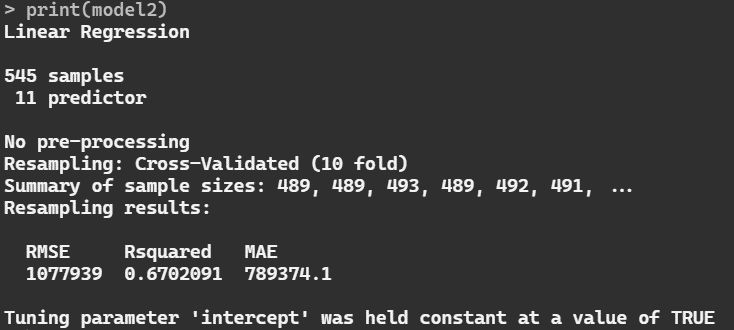


Figure 3.7: K fold cross validation results for full model and 11 predictors model in this order