Project Report

Introduction

This report investigates the best multiple linear regression model to predict median housing value of homes in California. The final model does not only indicate good predictions but also is simple enough to be understood by real estate agents, who are not familiar with technical terms. Therefore, the interpretability of the model is of high importance, and would be under consideration when any transformation is made to the data. According to real life, house price in the city is higher than in the rural area, and houses which location is convenient to reach essential service or transportation are also more expensive. Therefore, it is expected that house location will be one of the factors that influence median housing price of home in California.

Method

The model was built initially with all possible predictors, and the goal was to eliminate one by one predictor until we reached the best model. Before assessing the model validation, two conditions needed to be checked: conditional mean response is a single function of a linear combination of the predictors, and conditional mean of each predictor is a linear function with another predictor. If two conditions were satisfied, model violations could be assessed by residual plot. Otherwise, transformation on response or predictors or both was vital to satisfy mentioned conditions. Another significant problem should be brought to attention was multicollinearity among predictors. Multicollinearity could result in some problems with the model such as many predictors might be non-significant individually, but the overall F-test was highly significant, or the standard errors of the regression coefficients were much larger than they were supposed to be. Multicollinearity was checked by using Variance Inflation Factor. For those predictors whose VIF was greater than 5, further investigation involved using predictors as response was necessary. After that, the model would be respecified with less predictors to avoid multicollinearity. When looking at residual plot, three issues should be considered: linear relationship, error independence, and constant error variance. Normality of errors would be examined by QQ plot. Moreover, high leverage points, outliers, and influential points were also taken into account while building the model. The final best possible model would be built again in validation set to make sure that there would be minimal differences in estimated coefficients, which indicated the model was not overfitting.

Results

The original data consisting of 1000 observations was divided randomly into 2 data set: training data and validating data. Each of the data set contains 500 observations. We started by building a model with 13 predictors from the training data, and the summary of the model indicated that there exists a relationship between the median housing price and all the predictors in general, and also most of the predictors were individually linearly related to the response, only t-value of 2 predictors total_rooms and households did not show this relationship with the response. Before assessing model validity, I checked 2 conditions. It was clearly that in figure 1, for numerical variables, there were either linear relationship between predictors pairwisely or no relationship at all. Hence, no linear relationship violations occurred. Condition 2 was satisfied.

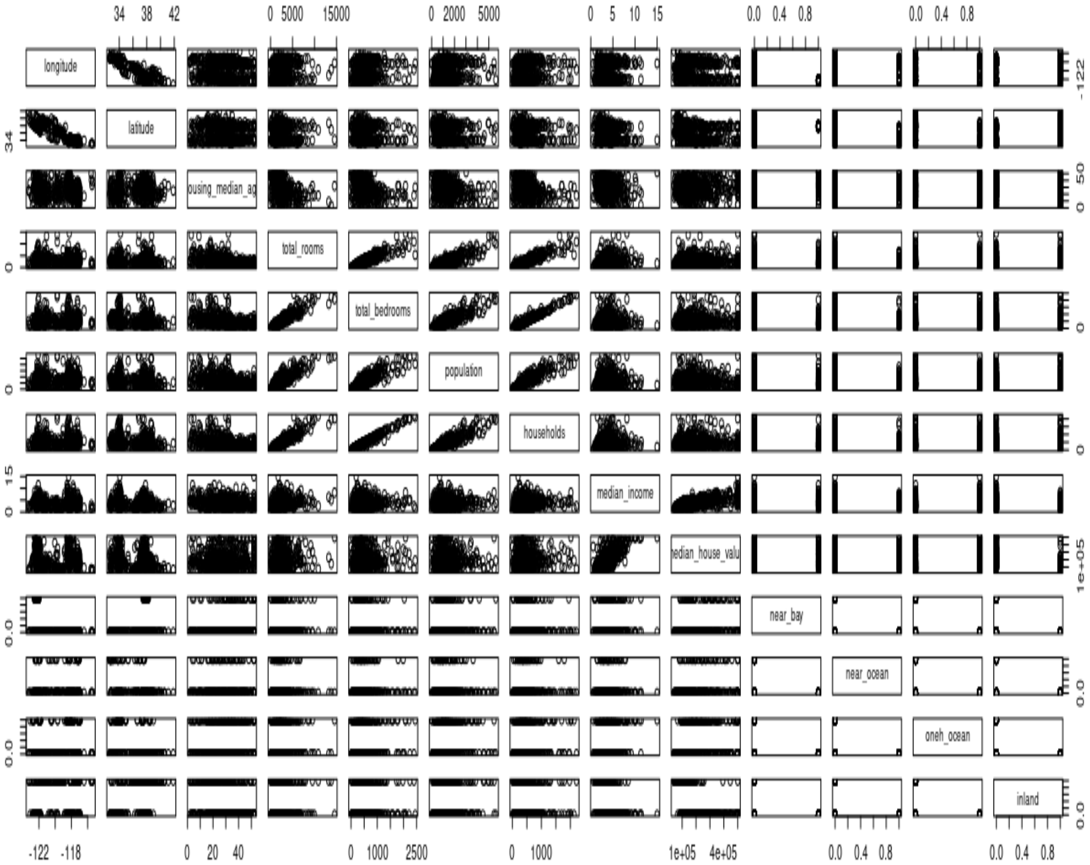


Figure 1: Relationship among all predictors

On the other hands, condition 1 was violated since there was curvature relationship between Y and Fitted Y. This required transformation method. I applied square root function on the response since in the graph between Y and Fitted Y, the curve assembles quadratic function. Refitting the model with new response =(median house price)$^{1/2}$ made the lowess line between Y and Fitted Y straighter, and the blue line was not a curve anymore. This is illustrated by Figure 2 below.

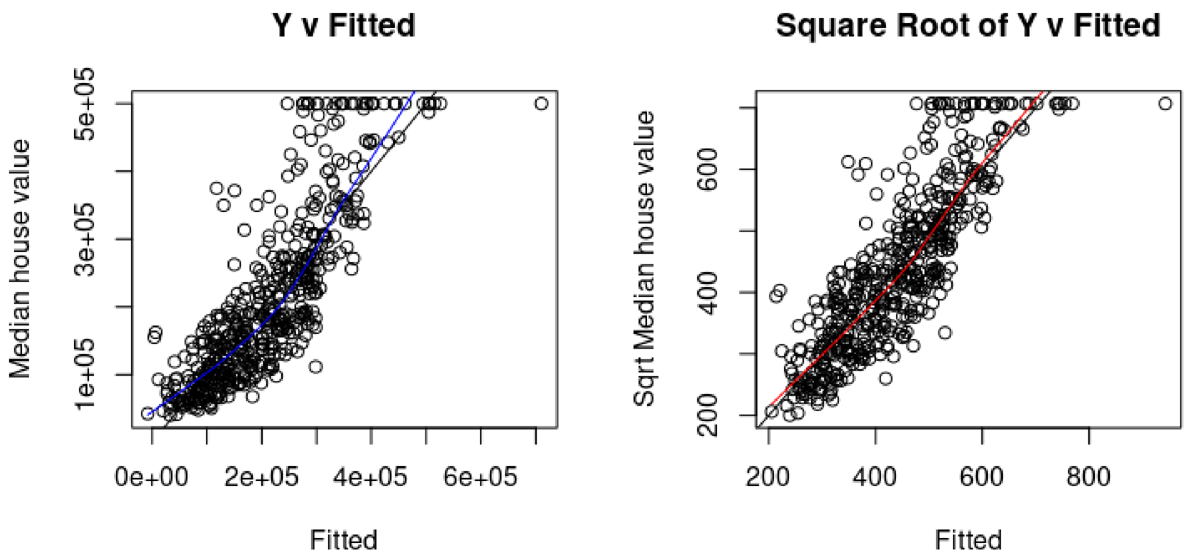


Figure 2: Before and After Transformation

Since the condition 1 was satisfied, I looked at residual plot. According to Figure 3, there existed outliers, non constant error variance, and the normality should also be of concern.

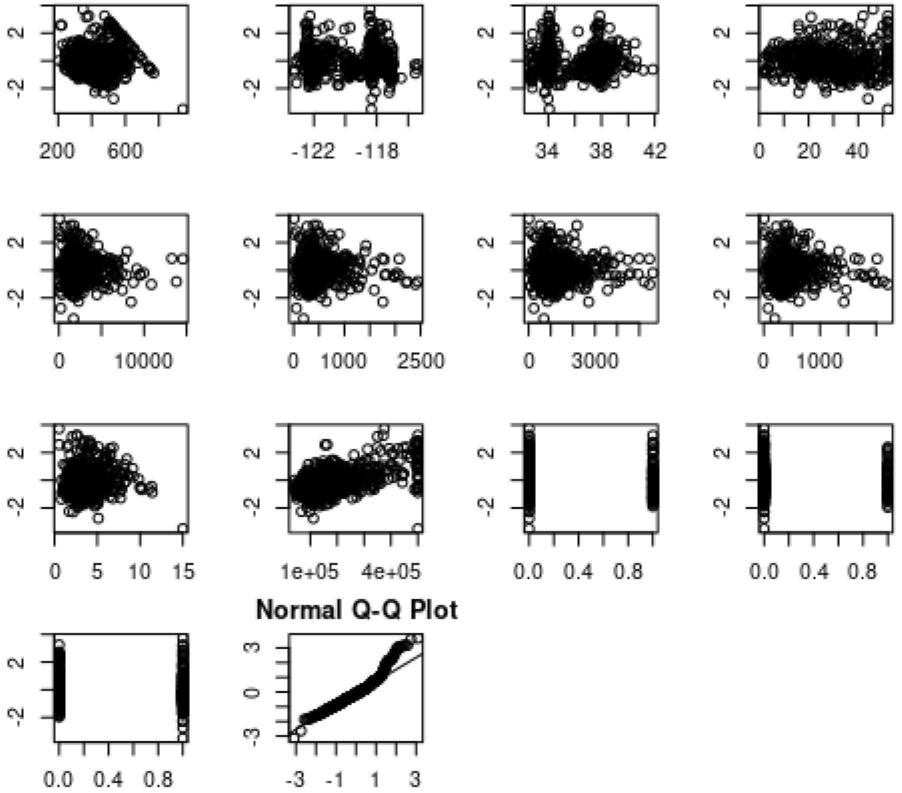


Figure 3: Residual Plot and QQ plot

However, before checking high leverage points, outliers, and influential points, I looked at multicollinearity since it might be one of the reasons causing this abnormal pattern. The predictors longitude, latitude, total_rooms, total_bedrooms, population, households, near_bay, near_ocean, onh_ocean, inland all had very high VIF. Since median income was the predictor having strongest relationship with response as indicated in model summary, I started to look at the linear model in which median income was the response, and with existence of other variables, total_rooms variable exhibited very strong relationship with median_income. Therefore, I decided to take it out of the model. The households predictor was also removed after checking model where population was the response, and inland was discarded after checking model where ocean was the response. I also attempted to remove oneh_ocean or near_bay since they were highly correlated to near_ocean; however, without them, the model performed worse.

The next step was to looking at points that might affect the model catastrophically. There were 29 leverage points, 28 outliers, and 3 out of 29 was bad leverage point. There was no influential point by Cook's Distance, 34 influential points by DFFITS, 57 influential points by DFBETAS. Refitting the model without bad leverage points and outliers gave us a model with adjusted R-squared equaled 0.825, and both t-test and F-test agreed about the strong linear relationship between predictors and response. To validate this model, I used validating data set consisted of 500 observations. The summary of this model gave adjusted R-squared equal 0.6987, which was a pretty good score, and the predictors as a whole or individually all exhibited strong linear relationship with the response, which was square root of median house price. The 2 tables below was the summary of the final models fitting on training and validation data set.

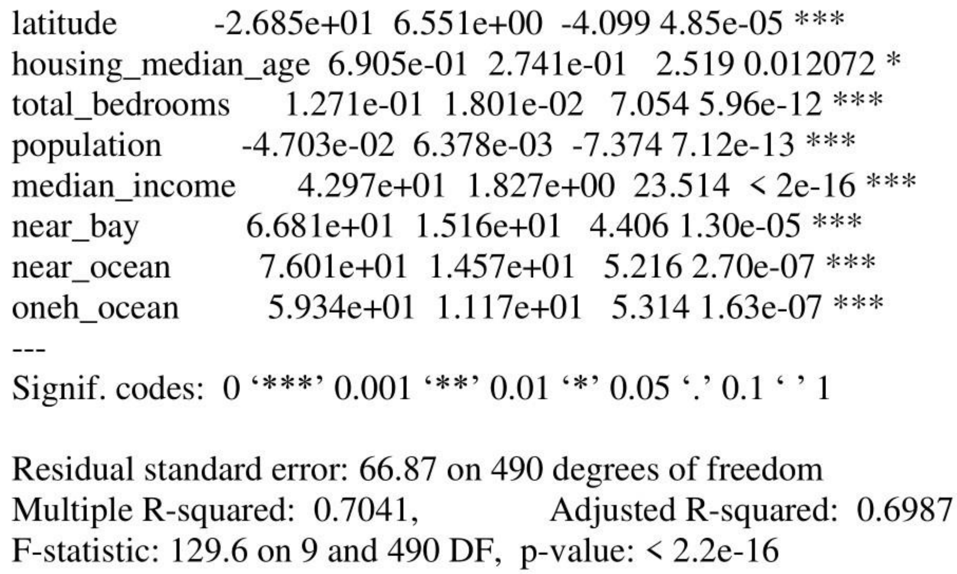
Table 1: Summary of Final Model

3. Final Model with training data

```
Call:
lm(formula = I(sqrt(median_house_value)) ~ ., data = traindata[-c(w1,
   w4, w5), -c(1, 5, 8, 14)])

Residuals:
    Min       1Q   Median       3Q      Max
-114.311  -35.896   -4.107   33.232  147.067

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)        -2.688e+03  5.478e+02  -4.906 1.33e-06 ***
longitude          -3.362e+01  6.492e+00  -5.179 3.46e-07 ***
latitude           -3.214e+01  6.472e+00  -4.966 9.93e-07 ***
housing_median_age  9.669e-01  2.499e-01   3.869 0.000126 ***
total_bedrooms      1.478e-01  1.574e-02   9.391  < 2e-16 ***
population         -5.041e-02  6.157e-03  -8.188 3.19e-15 ***
median_income       4.318e+01  1.500e+00  28.777  < 2e-16 ***
near_bay            4.521e+01  1.277e+01   3.540 0.000445 ***
near_ocean          6.083e+01  1.434e+01   4.242 2.73e-05 ***
oneh_ocean          5.956e+01  1.063e+01   5.605 3.76e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 51.11 on 422 degrees of freedom
Multiple R-squared:  0.8286,	Adjusted R-squared:  0.825
F-statistic: 226.7 on 9 and 422 DF,  p-value: < 2.2e-16
```

4. Final Model with validation data

```
Call:
lm(formula = I(sqrt(median_house_value)) ~ ., data = validdata[,
   -c(1, 5, 8, 14)])

Residuals:
    Min       1Q   Median       3Q      Max
-203.253  -41.732   -5.353   32.408  283.725

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)        -2.144e+03  5.630e+02  -3.809 0.000157 ***
longitude          -2.768e+01  6.635e+00  -4.171 3.58e-05 ***
```



```
latitude          -2.685e+01  6.551e+00  -4.099 4.85e-05 ***
housing_median_age  6.905e-01  2.741e-01   2.519 0.012072 *
total_bedrooms      1.271e-01  1.801e-02   7.054 5.96e-12 ***
population         -4.703e-02  6.378e-03  -7.374 7.12e-13 ***
median_income       4.297e+01  1.827e+00  23.514  < 2e-16 ***
near_bay            6.681e+01  1.516e+01   4.406 1.30e-05 ***
near_ocean          7.601e+01  1.457e+01   5.216 2.70e-07 ***
oneh_ocean          5.934e+01  1.117e+01   5.314 1.63e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 66.87 on 490 degrees of freedom
Multiple R-squared:  0.7041,	Adjusted R-squared:  0.6987
F-statistic: 129.6 on 9 and 490 DF,  p-value: < 2.2e-16
```

Discussion

The final model consists of predictors: longitude, latitude, housing_median_age, total_bedrooms, population, median_income, near_bay, near_ocean, oneh_ocean, and response is square root of median house price. In the training set, this model gave very high adjusted R-squared so there might be a concern about overfitting. However, in the validation data (which was a separate data set from the one that this model was built on), this model also performed fairly well with adjusted R-squared was approximately 0.7. This means the model can explain 70% of the variations in the data. However, this model might be improved further by transformation on predictor variables as well.

References

Prof. Daignault, Katherine. "Weblogin Idpz | University Of Toronto". *Q.Utoronto.Ca*, 2021, https://q.utoronto.ca/courses/204822.

This was my learning from STA302H1S – Winter 2021 at University of Toronto.