STAT 425

FINAL PROJECT

FALL 2021

TEAM 52

All coding and writing of sections by: Sagar Katiyar

INDEX

Section 1: Introduction

The goal of this project is to predict the house prices per unit area of houses in the Sindian Dist., New Taipei City, Taiwan using a combination of variables that are the transaction date, the house age, distance to nearest MRT station, the number of convenience stores in the living circle on foot, the geographical latitude and the geographical longitude. The data was provided by Prof. I-Cheng Yeh from the Department of Civil Engineering, Tamkang University, Taiwan.[1]

The project starts with an exploratory analysis of the data regarding its structure, type of variables, possible missing data and some plots to understand the relationships between the variables. Then in the analysis section of the data, two approaches are used. The first is a simple multiple linear regression model using variables using a backwards selection approach on which model diagnostics and model checking are done after which the model is tested and the prediction errors are calculated. The second model that is explored is a multivariate adaptive regression splines model and the results are compared to the results of the multiple linear regression model based on the prediction errors.

The project ends with a discussion of the results and the main findings of this project are summarized using a few bullet points.

Section 2: Exploratory Data Analysis

Including the variable x7 that is extracted from x1 the transaction date, there are a total of 9 columns in the provided dataset. They are as follows:

1. No: Representing the row number in numerical order
2. X1: Transaction date
3. X2: The age of the house
4. X3: The distance to the nearest MRT station
5. X4: The number of convenience stores in the living circle on foot
6. X5: The geographic coordinate, latitude
7. X6: The geographic coordinate, longitude
8. X7: The transaction month extracted from X1
9. Y: The house price of unit area, to be predicted from the above variables.

The structure of the data and a numerical summary of each variable can be seen below:

```
tibble [414 x 9] (S3: tbl_df/tbl/data.frame)
 $ No                                    : num [1:414] 1 2 3 4 5 6 7 8 9 10 ...
 $ X1 transaction date                   : num [1:414] 2013 2013 2014 2014 2013 ...
 $ X2 house age                          : num [1:414] 32 19.5 13.3 13.3 5 7.1 34.5 20.3 31.7 17.9 ...
 $ X3 distance to the nearest MRT station: num [1:414] 84.9 306.6 562 562 390.6 ...
 $ X4 number of convenience stores       : num [1:414] 10 9 5 5 5 3 7 6 1 3 ...
 $ X5 latitude                           : num [1:414] 25 25 25 25 25 ...
 $ X6 longitude                          : num [1:414] 122 122 122 122 122 ...
 $ Y house price of unit area            : num [1:414] 37.9 42.2 47.3 54.8 43.1 32.1 40.3 46.7 18.8 22.1 ...
 $ X7 transaction month                  : Factor w/ 12 levels "1","2","3","4",..: 12 12 8 7 11 9 9 6 7 6 ...
```

Figure 2.1: Structure of the dataset

```
       No        X1 transaction date  X2 house age    X3 distance to the nearest MRT station X4 number of convenience stores
 Min.   :  1.0   Min.   :2013        Min.   : 0.000   Min.   :  23.38                         Min.   : 0.000
 1st Qu.:104.2   1st Qu.:2013        1st Qu.: 9.025   1st Qu.: 289.32                         1st Qu.: 1.000
 Median :207.5   Median :2013        Median :16.100   Median : 492.23                         Median : 4.000
 Mean   :207.5   Mean   :2013        Mean   :17.713   Mean   :1083.89                         Mean   : 4.094
 3rd Qu.:310.8   3rd Qu.:2013        3rd Qu.:28.150   3rd Qu.:1454.28                         3rd Qu.: 6.000
 Max.   :414.0   Max.   :2014        Max.   :43.800   Max.   :6488.02                         Max.   :10.000

  X5 latitude     X6 longitude   Y house price of unit area X7 transaction month
 Min.   :24.93   Min.   :121.5   Min.   :  7.60             6      : 58
 1st Qu.:24.96   1st Qu.:121.5   1st Qu.: 27.70             7      : 47
 Median :24.97   Median :121.5   Median : 38.45             2      : 46
 Mean   :24.97   Mean   :121.5   Mean   : 37.98             12     : 38
 3rd Qu.:24.98   3rd Qu.:121.5   3rd Qu.: 46.60             4      : 32
 Max.   :25.01   Max.   :121.6   Max.   :117.50             11     : 31
                                                            (Other):162
```

Fig 2.2: A numerical summary of all variables

Of the above variables, No, X1, X2, X3, X4, X5, X6 and Y are quantitative variables whereas X7 is a categorical variables.

From the structure, it can be seen that all of the data is numeric data, with 9 columns and 414 rows.

Going with a 70-30 split of training and test data, 290 rows are allocated to the training data and 124 rows are allocated to the test data.

Some interesting points to note from the summary data:

1. Most of the transaction dates are from 2013 compared to 2014.
2. Similarly, most of the longitude values are the same as 121.5 compared to 121.6. There is much more variation in the latitude values ranging from 24.93 to 25.01. This could also be due to the degree of precision of the two variables.
3. No just represents the row numbers and transaction month is a categorical variable therefore the summary statistics for these variables do not convey any useful information.

The transaction month have been converted to factor variables since they are categorical in this dataset.

The following two plots show the pairwise correlation between the variables. The first plot shows the correlation between all variables whereas the second plot zooms in on the correlation between variables and the house price variable, Y, which is the one of interest in the exploration.

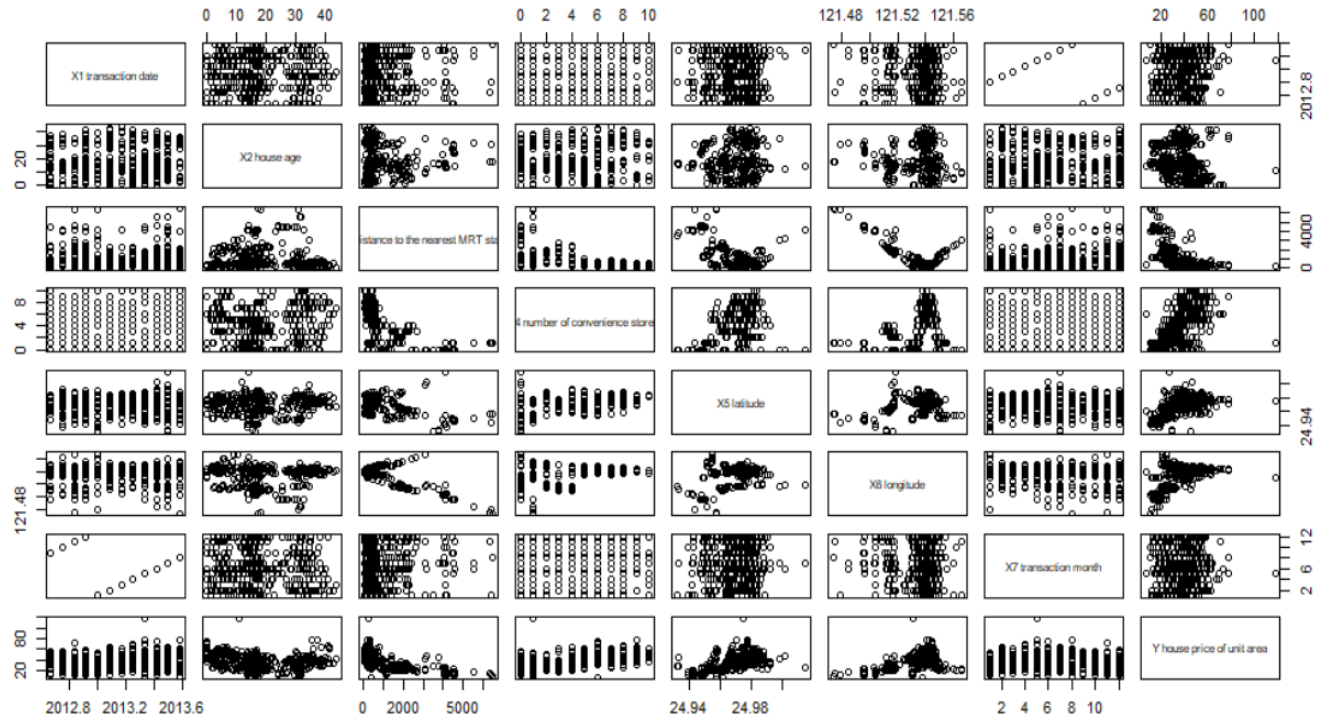

Fig 2.3: Correlation between variables X1-X7 and Y

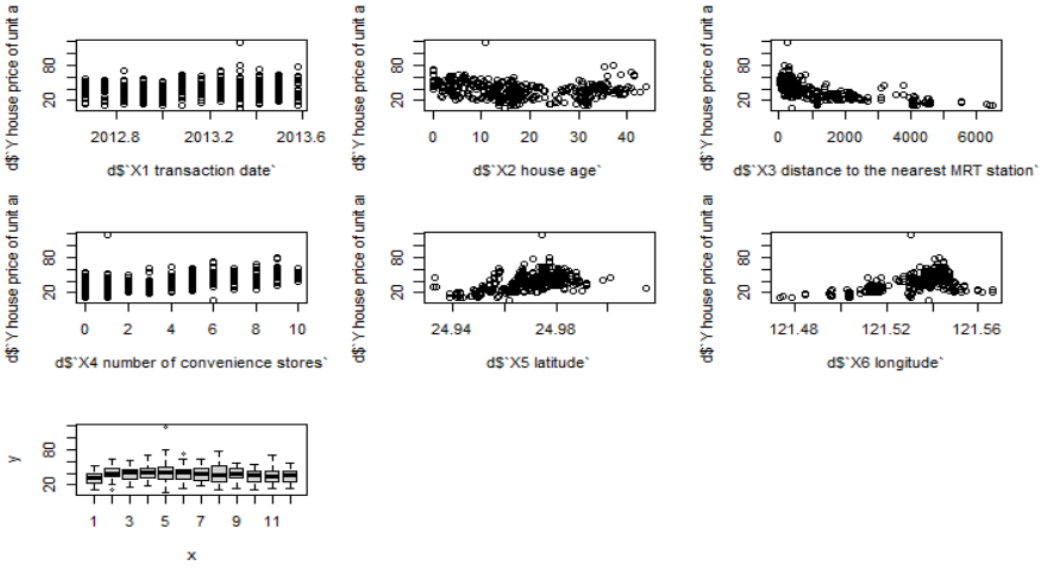

Fig 2.4: Correlation between variables and the house price, Y

While not all trends between variables are clearly visible, some clear relationships that can be seen are:

1. The distance to the nearest MRT station and the latitude and longitude. As the latitudes and longitudes increase to a certain point the distance to MRT stations decreases to a certain point and then starts increasing. This makes sense because the MRT stations since they might be located at specific coordinates.
2. A similar relationship can be seen between the number of convenience stores and the latitude and longitude which indicates that they are localized to certain areas.
3. As the distance to the nearest MRT station increases the house prices of units seem to drop which also makes sense since house prices would expect to go up the closer it is to the MRT stations.

4. There also appears to be a slight increase in average house prices as the number of convenience stores increases but it appears to be a weak correlation.
5. There appears to be a lack of relationship between house age and house prices which seems to be surprising since one would expect the prices to decrease as the age increases.
6. The relationship between house prices and latitudes/longitudes are similar to the relationship between the distance to the nearest MRT stations and latitudes/longitudes. The relationship between house prices and distance to the nearest MRT stations might be a possible reason for this relating correlation.

The above relationships were identified to be possibly important when fitting the prediction models and evaluating the conclusions. The next section explores two models that attempt to predict the house prices based on a combination of the other variables.

Section 3: Methods

After the train-test split of the data, the two models that are explored in this project are a simple multiple linear regression model with backwards selection for variable selection and a multivariate adaptive regression splines model.

Section 3.1: Multiple Linear Regression

The first fit model in this project was a simple multiple linear regression model with variables from X1 to X7 being used as predictors to predict Y. From this model, further variable selection was done using backwards selection and then model diagnostics and model checking methods were done on this model.

After fitting the multiple linear regression model and using backwards selection for selecting the variables, the resulting model had five variables including the intercept. These were the house age, distance to the nearest MRT station, number of convenience stores and the longitude. A full summary of the resulting model can be seen in table 3.1.1

Coefficients	Estimate	Pr(>|t|)	Multiple R-squared	p-value for model
Intercept	-6.94e03	7.98e-07	-	-
X2 House age	-3.01e-01	1.68e-09		-
X3 distance to nearest MRT station	-3.90e-03	6.34e-10	-	-
X4 number of convenience store	1.11e+00	5.10e-06	-	-
X5 latitude	2.80e+02	6.85e-07	-	-
Whole model	-	--	0.5682	<2.2e-16

Table 3.1.1: Summary of the final model

As can be seen in Table 3.11, the whole model as well as the individual predictors of the resulting model are far lesser than the 0.05 threshold implying, they are significant with a multiple R-squared value of 0.5682.

From the model diagnostics, several observations about the data and model were made:

1. Based on the analysis of leverages, there appear to be 2 outliers – 1 at row 271 of the training set and the other at row 221.
2. Based on the analysis of the Cooks distance, there are 2 highly influential points, one at row 271 and one at row 149.

Taking these into account, a new model should be fit without row 271 since it is both an outlier and highly influential. After doing that the new fit model is shown in table 3.1.2

Coefficients	Estimate	Pr(>|t|)	Multiple R-squared	p-value for model
Intercept	-6.70e03	5.51e-08	-	-
X2 House age	-2.90e-01	3.96e-11		-
X3 distance to nearest MRT station	-3.51-03	2.13e-10	-	-
X4 number of convenience store	1.33e+00	9.35e-10	-	-
X5 latitude	2.70e+02	4.60e-08	-	-
Whole model	-	--	0.6307	<2.2e-16

Table 3.1.2 New model after removing the highly influential outlier

After doing that, the estimates change and cause the multiple R-squared value to 0.6307 from 0.5682 which is a huge increase,

3. Using the Breusch-Pagan test for homoscedasticity it was discovered that the p-value was 0.7814 which means that we fail to reject the null hypothesis of homoscedasticity.
4. Using the Shapiro-Wilk's test for normality, it was discovered that the p-value was 1.14e-08 which implies we reject the null hypothesis assuming normality.

5. The slope for the regression lines fitted to the residuals after removing the effect of each variable one-at-a-time, is not the same as the regression coefficient for that variable in the full model therefore we can conclude that the assumption of the linear structure is not appropriate

Due to the violation of more than 1 assumption, it might be a better idea to find an alternate model to describe the relationship between variables.

Section 3.2: Testing the Multiple Linear Regression model

The next step was to test the multiple linear regression model on the test data.

The root-mean square prediction error turned out to be 8.26 which is high for the predicted variable and indicates that alternative models must be explored.

The next section therefore employs the use of another model, the multivariate adaptive regression splines to attempt to increase the R^2 and decrease the root-mean square prediction error.

Section 3.3: Multivariate adaptive regression splines

Due to the inadequacies of the simple multiple linear regression model, a multivariate adaptive regression splines (MARS) model was also explored. This allows us to include more variables since regression splines only allow the usage of one independent variable.

The variables selected for this model were the house age, distance to the nearest MRT station, number of convenience stores and the latitude, same as the one used for the multiple linear regression model.

We get a different set of coefficients corresponding to the new model and a massively increased R squared value of 0.7512 from 0.6307. Furthermore, the root mean square prediction error is also lower for the MARS model at 7.39 as compared to the 8.26 of the simple multiple linear regression (SMLR) model.

This seems to indicate that the MARS model performs better than the SMLR in predicting the house prices for this dataset, which is what was intended.

Section 4: Discussion of Results and Conclusion

In this project two widely different methods were employed in order to predict house prices from a combination of other variables, both with their sets of pros and cons.

- In the simple multiple linear regression model, an R^2 of 0.6307 and a prediction RMSE of 8.26 were achieved whereas in the multiple adaptive regression splines model an R^2 of 0.7512 and a prediction RMSE of 7.39 were achieved.
- Purely in terms of results alone the latter performed significantly better on the test data therefore would be my preferred choice when faced with more data from the same environment.
- Moreover, it was clearly seen that some of the assumptions of the simple multiple linear regression model were violated, making the model prone to suspicion as the "correct" model.
- However, when it comes to interpretive understanding of the coefficients, the coefficients from the simpler multiple linear regression model are definitely easier to interpret. The coefficients of house age and distance to the MRT station are negative, which is to be expected since older properties tend to depreciate in price over time and the closer a house is to an MRT station, the higher the price should be. Moreover, the coefficient for the number of convenience stores is positive which also makes sense since it increases the potential of the house. On the other hand, the coefficients of the MARS model are not easily interpretable.

- Moreover, there is a constraint on the MARS model to only four variables therefore more complex relationships cannot be explored even if they might lead to a more accurate model. This restraint to the number of variables is not present in the multiple linear regression model.

Therefore, even though the MARS model provides more promising results and the multiple linear regression provided more flexibility and easier interpretability, I strongly believe that none of them are too close to the best model that can be created. A possible exploration into alternative models like non parametric regression and random forests might also potentially provide promising avenues to go down to predict the targeted variable. I also believe that the models that have been explored so far were promising enough to conclude that it is definitely possible to predict the house prices with a fairly high level of certainty given an exploration of a few more models.

Section 5: Acknowledgements

All of the knowledge and information that I acquired in order to do this project can be attributed to three sources. The first and biggest source of information was the course website through which I learnt how to do the exploratory data analysis, multiple linear regression, model diagnostics and testing. The second source was the UCI repository which hosted the dataset and background information pertaining to it. The third and final resource was a short YouTube video titled 'Regression in R: Simple Multivariate, Step-wise, Spline, MARS, and Loess' by Dr. Mondesire I referred to on the multiple additive regression splines model and its usage.