Predicting Housing Prices

By Zahra Aminiranjbar and Aditya Jhunjhunwala
Link to presentation slides: https://tinyurl.com/STA232-Group1

Introduction

The market historical data set of real estate valuation from Sindian Dist., New Taipei City, Taiwan is given. The response variable of interest is the house price per unit area, Y (unit: 10000 NT$/ping). The data is available for houses at different temporal locations in the city i.e. at different longitude and latitude. Additionally, the age of the house (referred to as age hereafter) and the date of the purchase are given (referred to as date hereafter). From the geographic information the following data is also obtained:

1) Distance to nearest MRT station - in meters (referred to as MRT)
2) No of convenience stores in the living circle on foot (integer) (referred to as stores)

The given data set has no missing data and a total of 414 house data. The date of purchases are from the 8th month of 2012 - 2012.667 and go up to the 7th month of 2013 - 2013.583. The 8th month of 2012 is set to be the start of the timeline so is set to 0. The housing price, age, MRT, and stores are plotted on the geographical location of the houses in Figure 1.

The objective of the project is to find the parameters on which the housing prices depend and a predictor model which can estimate the housing price given a location and the other predictor variables.

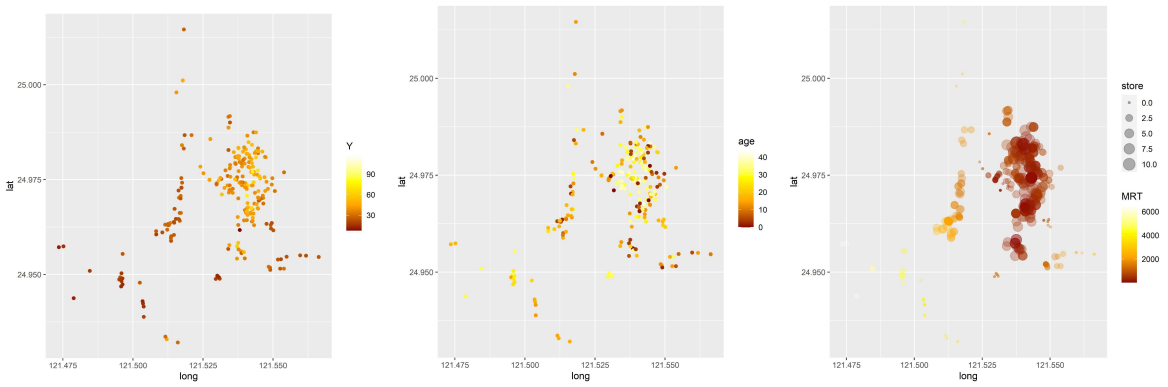

Figure 1: Scatter plot of house price, age of the house, no of stores, and distance to MRT at house locations

Correlation and Multicollinearity

The scatter plot of the variables, histogram of the 6 predictor variables, and the correlation matrix are shown in Figure 2. The following observations can be made:

- MRT displays a high negative correlation with Prices which is expected as the houses near a transportation hub will be costly
- No. of stores displays high positive correlation with Prices which is again expected.
- MRT and stores show high collinearity.
- Both date and age show no collinearity with other variables.
- Collinearity of Latitude and Longitude with MRT shows that houses are clustered near the MRT stations.
- MRT data is skewed to the left. This can be attributed to more data being collected for houses near the MRT stations or that the housing density is more near the MRT stations.

- Housing prices have outliers i.e. a few houses have extremely high prices.

	date	age	MRT	store	lat	long	Y
date	1	0.0175	0.0609	0.0095	0.035	-0.0411	0.0875
age	0.0175	1	0.0256	0.0496	0.0544	-0.0485	-0.2106
MRT	0.0609	0.0256	1	-0.6025	-0.5911	-0.8063	-0.6736
store	0.0095	0.0496	-0.6025	1	0.4441	0.4491	0.571
lat	0.035	0.0544	-0.5911	0.4441	1	0.4129	0.5463
long	-0.0411	-0.0485	-0.8063	0.4491	0.4129	1	0.5233
Y	0.0875	-0.2106	-0.6736	0.571	0.5463	0.5233	1

Figure 2: correlation matrix and histogram of housing cost (Y) (For scatter plot and histogram refer slides)

To summarise, the variables show a high correlation with each other. Hence, the linear models used would encounter issues of multicollinearity.

Model selection

As a baseline, we chose a naive model that uses all the raw data without any transformation. The data was then fitted with a linear regression model (our Full model).

$Y = \beta0 + \beta1*date + \beta2*age + \beta3*MRT + \beta4*store + \beta5*lat + \beta6*long$

The R squared calculated from the full model is 0.5824 which indicates that around 58% variability in the housing prices(Y) can be explained by our model. In order to better understand the performance of our model, we focused our attention on the plots of residuals versus fitted values and the variables. We observed violations of constant variance in residuals versus housing prices (Y) and distance to the nearest MRT (MR), Figure 3 . Furthermore, to test our null hypothesis, formulated as : H0: $\beta1=\beta2=\beta3=\beta4=\beta5=\beta6$ vs Ha: $\beta i \neq 0$ for i=1,..,6

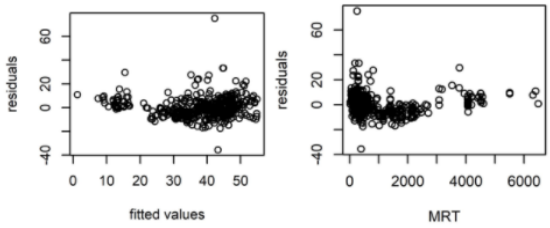

Figure 3: Residual vs fitted value and MRT

(were βs are the coefficients of the linear model) family-wise Bonferroni confidence intervals of the coefficients were calculated. Among them, the variable longitude confidence interval includes zero which indicates that we can not reject the null hypothesis for this variable. BIC/AIC criteria also had the lowest value for a model where all the variables except the longitude were present. Using the analysis above we came up with a new model (our submodel) that could better describe our data. The new model is:

$\log(Y) = \beta0 + \beta1*date + \beta2*age + \beta3*\log(MRT) + \beta4*store + \beta5*lat$

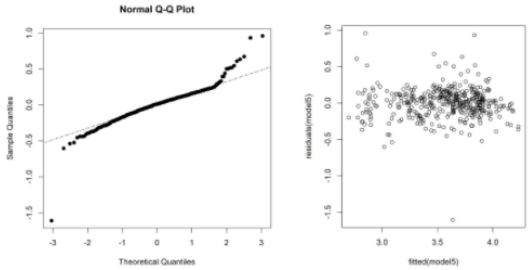

Figure 4: QQ plot and residual vs fitted value of submodel

The value of R squared improved from 0.5824 to 0.7181 and the residual plots of the fitted values demonstrated random distribution around zero, figure 4. The QQ plot demonstrates the normal distribution of the residuals and the residuals versus fitted value shows random distribution around zero. A summary of thevarious models tested (p-values of F-statistic & t-statistic, SS of the variables, R2, R2 adj) is shown in the summary table1.

Ridge regression

Whenever a model predicts there are prediction errors known as bias and variance. There is a tradeoff between bias and variance. A model with high variance performs well on the training data but has high error rates on the test data. On the other hand, a model with high bias under fits the data. In order to have a model with low variance and low bias, we used ridge regression on our submodel. We split the data into training and test set (80/20) and the optimum regularization parameter was chosen using 10 fold cross-validation. The value of R squared increased from 0.7181 to 0.7270. The increase is not that significant which we could already predict since the difference between the RMSE calculated on the training and test data was not significant. Therefore our submodel was not overfitting or underfitting.

			Intercept	date	age	MRT	store	lat	long	SSR	SSE	R2	R2 adj
model1	Y = date+age+MRT+store+lat+long	SS		585.90	3440.88	34857	3576.11	2065	5.13	44530	31931	0.5824	0.5762
		Pr(>|t|)	0.51	1.00E-03	1.10E-11	1.00E-09	3.80E-09	6.40E-07	0.80	P value for F:		2.20E-16	
model2	Y = date+age+ln(MRT)+store+lat long	SS		585.90	3440.88	41202	744.55	3904	18.79	49897	26564	0.6526	0.6475
		Pr(>|t|)	0.03	6.10E-06	2.20E-10	2.00E-16	0.06	3.50E-13	0.59	P value for F:		2.20E-16	
model3	Y = date+age+ln(MRT)+store+lat	SS		585.90	3440.88	41202	744.55	3904		49878	26583	0.6523	0.6481
		Pr(>|t|)	1.50E-13	5.70E-06	2.00E-10	2.00E-16	0.06	7.90E-14		P value for F:		2.20E-16	
model4	log(Y) = date+age+MRT+store+lat	SS		0.3626	2.3113	36.1551	2.2988	2.4928		43.6206	19.9956	0.6857	0.6818
		Pr(>|t|)	1.10E-11	5.30E-04	2.10E-12	2.00E-16	8.00E-09	4.50E-12		P value for F:		2.20E-16	
model5	**log(Y) = date+age+ln(MRT)+store+lat**	SS		0.3626	2.3113	36.8442	0.7703	5.3925		45.6809	17.9353	0.7181	0.7146
		Pr(>|t|)	2.00E-16	8.20E-06	1.20E-10	2.00E-16	3.90E-02	2.00E-16		P value for F:		2.20E-16	

Table 1: Statistic summary of various models tested.

Principal Component Regression (PCR)

The high multicollinearity of the data calls for Principal component (PC) regression which fits the regression model through an orthogonal transformation of the columns of the design matrix. The transformed basis vectors or principal components (PCs) are mutually orthogonal thereby solving the issue of multicollinearity. The vectors of the variables in the PC1 and PC2 space are shown in Figure 5(a). MRT, stores, Lat, and Long contribute primarily to the PC1 and age and date to PC2. The former set of variables was found to be highly correlated in the correlation matrix. The contribution of different PC's to the complete space of X'X i.e. the participation factor is shown in Figure 5(b). The plot of house prices (Y) with components of data projected on the PC1, hereby called PC1, with the PCs obtained from data with and without log transformation, as discussed in the previous section, is shown in Figure 6. It is clearly evident that the log-transformed model is suitable even for the PCR.

On fitting a linear model using the PCs of the not transformed data gave the same R2 as the full model - 0.5824. However, on fitting the data to the PCs of the log-transformed data (log(Y) and log(MRT)), an improved R2 of 0.724 was obtained compared to the R2 of 0.7181. The p-value for t-statistic for the PC2 and PC5 was 0.09 and 0.40 respectively. On fitting a submodel with PC1, PC3, PC4, and PC6, and R2 of 0.7215 were obtained. The p-value for the F statistic for Hypothesis that coefficients for PC2 and PC5 are zero was 0.1694 > 0.05. Thus, PC2 and PC5 can be neglected. The regression stats and estimates of the coefficients obtained from the submodel are shown in Figure 7.

A fit on randomly selected 80% of the dataset was used to fit the PCR with the 4 PCs (discussed above) of the log-transformed data and data without any transformation. The obtained fit was used to predict Y for the remaining 20% dataset. The predicted values and the actual values for the two submodels are shown in Figure 7. The prediction in the no transformation submodel shows not so random scatter around the 45° line. It is clearly evident that the log-transformed data predicts the values more robustly.

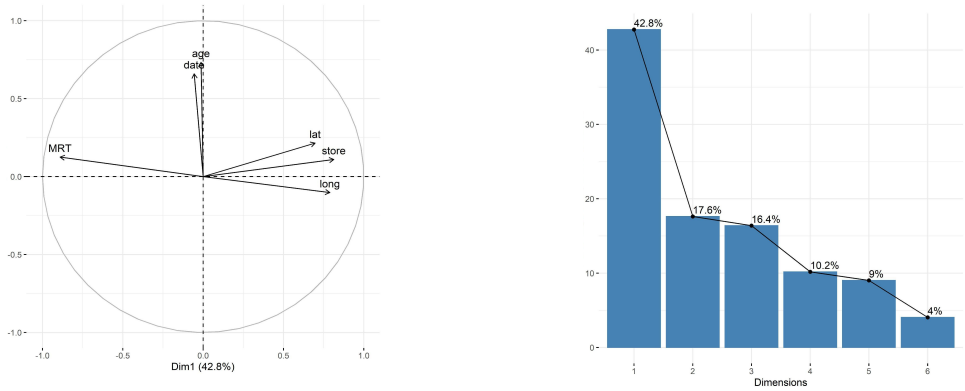

Figure 5: (a) Vector map of variable in PC1 and PC2 space. (b) Participation factor for the PC's

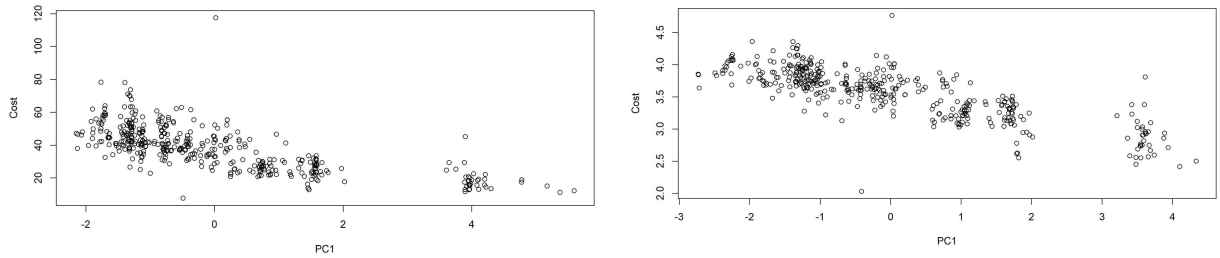

Figure 6: PC1 vs Y for (a) transformed data (b) not transformed data

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.566695   0.010228 348.703  < 2e-16 ***
PC1         -0.196994   0.006394 -30.810  < 2e-16 ***
PC3          0.083224   0.010329   8.057 8.64e-15 ***
PC4          0.045338   0.013107   3.459 0.000599 ***
PC6         -0.120480   0.020788  -5.796 1.36e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2081 on 409 degrees of freedom
Multiple R-squared:  0.7215,    Adjusted R-squared:  0.7188
F-statistic: 264.9 on 4 and 409 DF,  p-value: < 2.2e-16
```

Figure 7: Summary statistics for the submodel with four principal components of log-transformed data

Decision Tree

The decision tree algorithm works by partitioning the data into sub-spaces repeatedly until the outcome in each subspace is as homogenous as possible. Metrics such as residuals or mean square error(MSE) can be used to evaluate the quality of the regression tree. Figure 8 shows the best tree chosen using the decision tree algorithm. It is important to note that the root node in the decision tree is the node that

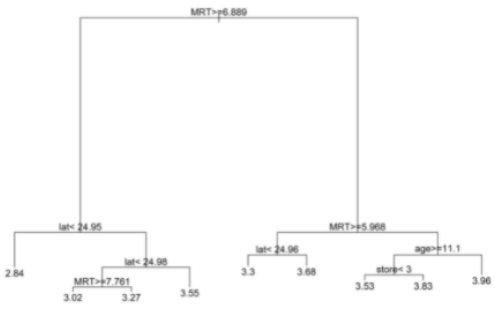

Figure 8: Best decision tree

best splits the data and here that variable is MRT. This indicates that the distance to the nearest MRT most explains the housing prices in our dataset. The prediction that our regression tree makes at the end is the average of the values for the target variable at such leaf nodes. The value of R Squared calculated from the decision tree is 0.6820. This value is smaller than the value of R squared calculated using the linear submodel 0.7181. Since the decision tree is prone to overfitting the training data set and it's sensitive to outliers. A single decision tree normally does not make great predictions, so multiple trees are often combined to make a forest to give birth to stronger ensemble models, such as a random forest. A random forest acts as an estimator algorithm that aggregates the result of many decision trees and then outputs the most optimal result. The value of R squared calculated from the random forest is 0.8076 which is a significant improvement compared to our best R squared value calculated so far.

Conclusion

The purpose of this project was to do data analysis on real estate housing prices and to come up with an innovative approach that can best model the data and explain the variability in our response variable Y (housing price). Our findings of the data visualization and model selections are as follows:

1. There is multicollinearity among the variables in the data set and the most important predictor of the housing prices is the distance to the nearest MRT. This is derived from the high correlation and also the root node of the decision tree.
2. Our model significantly improved with log transformation of both the response variable Y and the predictor variable distance to the nearest MRT.
3. An exhaustive search on model selection based on AIC/BIC criteria and as well as simultaneous Bonferroni confidence intervals suggests a simpler model with dropping longitude from the variables. The value of R squared was improved from 0.58 to 0.71.
4. Due to the presence of multicollinearity between predictors' latitude and distance to MRT, we carried out the principal component analysis. R squared calculated from the PCA using log transformation of Y and MRT, improved even when 4 PCS were used.
5. Our random forest model was able to describe 80% of the variation in housing prices compared to the full model of 58%, which is an improvement of 22%. Given our data sets, 80% can be considered a good value since housing prices are usually influenced by features such as the lot size, number of bedrooms, distance to highway, school, and crime rate in the area, which were not available in this dataset.

Participation

All the group members participated equally in the project.

References

- Yeh, I-Cheng, and Tzu-Kuang Hsu. "Building real estate valuation models with comparative approach through case-based reasoning." *Applied Soft Computing* 65 (2018): 260-271.
- Gareth, James, et al. *An introduction to statistical learning: with applications in R*. Spinger, 2013.
- http://www.sthda.com
- https://towardsdatascience.com